Computer Science Department

TECHNICAL REPORT

THE FORMULATION AND ANALYSIS OF NUMERICAL METHODS FOR INVERSE EIGENVALUE PROBLEMS

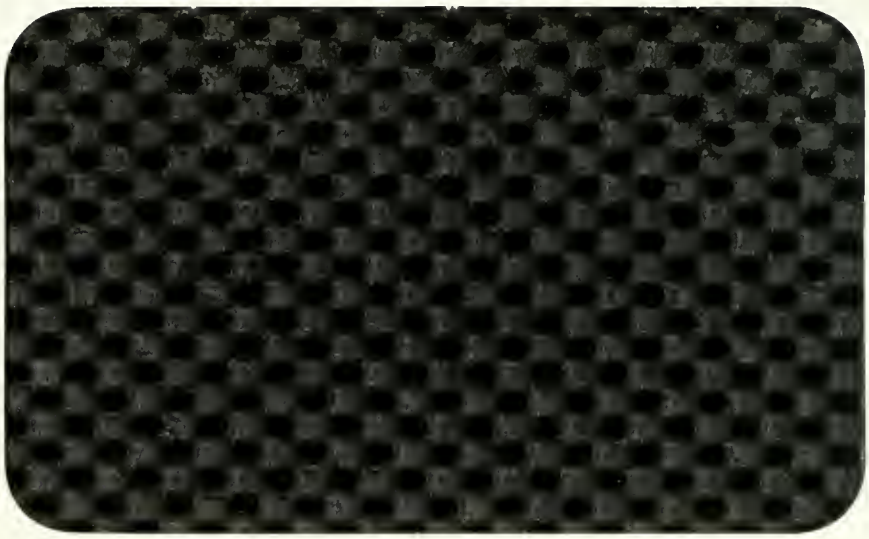
By

S. Friedland[1]
J. Nocedal[2]
M. Overton[3]

September 1985
Technical Report #179

NEW YORK UNIVERSITY

Department of Computer Science
Courant Institute of Mathematical Sciences
251 MERCER STREET, NEW YORK, N.Y. 10012



THE FORMULATION AND ANALYSIS OF NUMERICAL METHODS FOR INVERSE EIGENVALUE PROBLEMS

By

S. Friedland[1]
J. Nocedal[2]
M. Overton[3]

September 1985
Technical Report #179

ABSTRACT. We consider the formulation and local analysis of various quadratically convergent methods for solving the symmetric matrix inverse eigenvalue problem. One of these methods is new. We study the case where multiple eigenvalues are given: we show how to state the problem so that it is not overdetermined, and describe how to modify the numerical methods to retain quadratic convergence on the modified problem. We give a general convergence analysis which covers both the distinct and the multiple eigenvalue cases. We also present numerical experiments which illustrate our results.

---

[1] Institute of Mathematics, Hebrew University , and Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago Il 60680. The work of this author was supported in part by the National Science Foundation Grant No. MCS 83-00842

[2] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Il 60201. The work of this author was supported in part by the National Science Foundation Grant No. DCR-84-01903.

[3] Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York N.Y. 10012. The work of this author was supported in part by the National Science Foundation grants No. DCR-83-02021 and DCR-85-02014

THE FORMULATION AND ANALYSIS OF NUMERICAL METHODS FOR INVERSE EIGENVALUE PROBLEMS

S. FRIEDLAND[1], J. NOCEDAL[2], AND M.L. OVERTON[3]

ABSTRACT. We consider the formulation and local analysis of various quadratically convergent methods for solving the symmetric matrix inverse eigenvalue problem. One of these methods is new. We study the case where multiple eigenvalues are given: we show how to state the problem so that it is not overdetermined, and describe how to modify the numerical methods to retain quadratic convergence on the modified problem. We give a general convergence analysis which covers both the distinct and the multiple eigenvalue cases. We also present numerical experiments which illustrate our results.

[1]Institute of Mathematics, Hebrew University, and Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago Il 60680. The work of this author was supported in part by the National Science Foundation Grant No. MCS 83-00842

[2]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Il 60201. The work of this author was supported in part by the National Science Foundation Grant No. DCR-84-01903.

[3]Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York N.Y. 10012. The work of this author was supported in part by the National Science Foundation grants No. DCR-83-02021 and DCR-85-02014 .

§1. INTRODUCTION

Let $A(c)$ be the affine family

$$A(c) = A_0 + \sum_{k=1}^{n} c_k A_k, \tag{1.1}$$

where $c \in \mathbf{R}^n$ and $\{A_k\}$ are real symmetric nxn matrices. Denote the eigenvalues of $A(c)$ by $\{\lambda_i(c)\}_1^n$, where

$$\lambda_1(c) \leq ... \leq \lambda_n(c).$$

The following is called an *inverse eigenvalue problem*:

PROBLEM 1. *Given real numbers* $\lambda_1^* \leq ... \leq \lambda_n^*$, *find* $c \in \mathbf{R}^n$ *such that* $\lambda_i(c) = \lambda_i^*$, $i = 1, ..., n$.

There is a large literature on conditions for existence and uniqueness of solutions to Problem 1 (or its variations) in many special cases. In this paper we are concerned with the formulation and local analysis of various quadratically convergent methods to solve the problem, assuming the existence of a solution. Extending our techniques to give methods with good global behavior is an important task which we shall not explicitly address.

The paper is organized as follows. In §1.1 we discuss some of the important motivating applications which arise in the physical and social sciences. Many of these lead to closely related variations of the model problem given above. In §2 we confine our attention to the case where the given eigenvalues $\{\lambda_i^*\}_1^n$ are distinct, and describe several numerical methods. Four of these are related to Newton's method and are generally locally quadratically convergent. Of these four methods, three are known in the literature, and one is apparently new. In §3 we discuss the case where multiple eigenvalues are present in the set $\{\lambda_i^*\}_1^n$. It is well known that the eigenvalues are not differentiable functions at the points where they coalesce. Nonetheless, the behavior of the numerical methods in these circumstances has received little attention. In §3.1 we discuss the case where the numerical methods of §2 are applied, without modifications, to problems with multiple eigenvalues. Assuming Problem 1 has a solution, we show that the methods retain local quadratic convergence, with

little or no modification, even though the eigenvalues are not differentiable at the solution. In §3.2 we argue that Problem 1 is generally overdetermined when multiple eigenvalues are present, and show how to modify the problem so that it has the appropriate number of parameters and target eigenvalues. We then explain how to modify the numerical methods of §2 to retain quadratic convergence on the modified problem. In §3.3 we give a general convergence analysis which covers both the distinct and the multiple eigenvalue cases. In §4 we present numerical experiments which illustrate our results.

Before we proceed we must mention that in many applications the problem to be solved is different from Problem 1. Sometimes $A(c)$ is a nonlinear matrix function of c. In §2.1 we briefly discuss how to adapt the numerical methods for this case. Other applications lead to variations of Problem 1 that include the following: the number of given eigenvalues is less than n, the order of the matrices; the number of parameters is not the same as n; there are constraints on c; there is a functional to be minimized subject to eigenvalue constraints of the form given by Problem 1; the constraints on some of the eigenvalues are inequalities instead of equalities. (This last case seems to be particularly common in practical applications.) In §1.1 we give a few examples to illustrate how some of these applications arise. We think that it is not difficult to see how the problem formulations, numerical methods, and convergence analyses can be extended to some of the variations of Problem 1. However we shall not give any details here.

A special case of Problem 1, which is frequently encountered, is obtained when the linear family (1.1) is defined by

$$A_k = e_k e_k^T \qquad k = 1, ..., n,$$

where e_k is the k^{th} unit vector, so that

$$A(c) = A_0 + D \tag{1.2}$$

where $D = diag(c_k)$. This problem is known as the *additive inverse eigenvalue problem.* Conditions for existence and uniqueness of solutions to this problem are well understood.

Friedland (1977) showed that the problem is always solvable over the complex field, and it is easy to construct examples that show that it is not always solvable over the reals.

§1.1 Applications.

We will now describe several inverse eigenvalue problems arising in various areas of application.

One classical example is the solution of inverse Sturm-Liouville problems. Consider for example the boundary value problem

$$-u''(x) + p(x)u(x) = \lambda u(x)$$
$$u(0) = u(\pi) = 0.$$

Suppose that the potential $p(x)$ is unknown, but the spectrum $\{\lambda_i^*\}_1^\infty$ is given. Can we determine $p(x)$? This continuous problem has been studied by many authors; see Borg (1946), Gelfand and Levitan (1955) and Hald (1972). We are mainly interested in the discrete analogue of this problem. Let us use a uniform mesh, defining $h = \frac{\pi}{n+1}$, $u_k = u(kh)$, $p_k = p(kh)$, $k = 1, \ldots, n$, and assume that $\{\lambda_i^*\}_1^n$ is given. Using finite differences to approximate u'' we obtain

$$\frac{-u_{k+1} + 2u_k - u_{k-1}}{h^2} + p_k u_k = \lambda_j^* u_k, \qquad k = 1, \ldots, n \qquad u_0 = u_{n+1} = 0, \tag{1.3}$$

where λ_j^* is an eigenvalue in the set $\{\lambda_i^*\}_1^n$. Thus we have an additive inverse eigenvalue problem (1.2) with

$$A_0 = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & & \ddots & \\ & & & & 2 \end{pmatrix} \tag{1.4}$$

and $D = diag(p_k)$. Hald (1972) is a comprehensive reference for both the continuous and discrete inverse Sturm-Liouville problem.

Another interesting inverse eigenvalue problem is obtained by studying a vibrating string. Here the boundary value problem is

$$\begin{aligned} -u''(x) &= \lambda \rho(x) u(x) \\ u(0) &= u(\pi) = 0. \end{aligned} \tag{1.5}$$

The question is whether we can determine the density function $\rho(x) > 0$ from the eigenvalues $\{\lambda_i^*\}_1^\infty$. Discretizing the problem as before we obtain

$$Au = \lambda_i^* Du \qquad i = 1, \ldots, n$$

or equivalently

$$D^{-1}Au = \lambda_i^* u \qquad i = 1, \ldots, n, \tag{1.6}$$

where $D = diag\big(\rho(kh)\big) > 0$ and A is given by the right hand side of (1.4). This kind of problem is called the *multiplicative inverse eigenvalue problem*: given a real symmetric matrix A and eigenvalues $\{\lambda_i^*\}_1^n$, find a positive diagonal matrix V such that VA has the given eigenvalues. We can write this problem in the form (1.1) where $A_0 = 0, A_k = e_k a_k^T, k = 1, \ldots, n$, and where a_k^T is the k^{th} row of A. The matrices A_k are not symmetric in this case. Note, however, that a diagonal similarity transformation applied to VA gives the symmetric matrix $V^{\frac{1}{2}} A V^{\frac{1}{2}}$.

Both the additive and multiplicative inverse eigenvalue problems were posed by Downing and Householder (1956). In practical applications of the inverse Sturm-Liouville and inverse vibrating string problems, only a few of the smallest eigenvalues may be given. In order for the problem to be well-posed, the number of parameters must be reduced accordingly. This can be done by expressing the potential or density function as a linear combination of a few given basis functions. See Osborne (1971) and Hald (1972) for details.

Problem 1 also arises in nuclear spectroscopy (see Brussard and Glaudemans (1977)). There $A(c)$ is the Hamiltonian and the set $\{\lambda_i^*\}$ is obtained from experimental measurements. A similar problem occurs in molecular spectroscopy (see Pliva and Toman (1966) and Friedland (1979)). Practical formulation of these problems often involves a number of

parameters which is smaller than the number of given eigenvalues. It is therefore appropriate to consider a least squares formulation:

$$\min_{c \in \mathbf{R}^n} \sum_{i=1}^{m} \left(\lambda_i(c) - \lambda_i^*\right)^2. \tag{1.7}$$

The methods that we shall discuss for solving (1.1) can be generalized to handle (1.7) by using well known techniques (see, for example, Dennis and Schnabel (1983), chapter 10).

The *problem of communality*, which arises in factor analysis (see Harman (1967), chapter 5), is as follows. Let A_0 be a given real symmetric matrix with zero diagonal entries. The objective is to find a diagonal matrix D such that $A_0 + D$ has minimal rank. In other words, the goal is to find D such that $A_0 + D$ has as many eigenvalues equal to zero as possible. This problem is different from Problem 1, since neither the rank nor the nonzero eigenvalues are known. However, we can guess the rank and hence the number of eigenvalues which are equal to zero. In some cases this is enough to locally determine a solution, as we shall explain in §3.

In the *educational testing problem* (see Fletcher(1985)), we are given a symmetric positive definite matrix A_0 and want to know how much can be subtracted from the diagonal of A_0, with the restriction that the resulting matrix is positive semi-definite. The problem may be posed as follows:

$$\begin{aligned} &\max_{c} \sum_{k=1}^{n} c_k \\ \text{subject to} \quad &\lambda_i(A_0 - D) \geq 0 \qquad i = 1, \ldots, n \\ &D = diag(c_k) \geq 0. \end{aligned} \tag{1.8}$$

In this problem, as in the problem of communality, we can usually expect a multiple zero eigenvalue at the solution. Fletcher (1985) also describes a problem that has the same structure as (1.8), which he calls the *matrix modification problem*. We are given a symmetric indefinite matrix A_0 and want to add as little as possible to the diagonal of A_0 to obtain a positive semi-definite matrix.

An important class of problems frequently occurring in engineering applications has the form

$$\begin{aligned} &\min_{c \in \mathbb{R}^m} f(c) \\ &\text{subject to} \quad l \leq \lambda_i(c) \leq u \quad i = 1, \ldots, n, \end{aligned} \tag{1.9}$$

where $f(c)$ is a real-valued objective function and l and u are specified lower and upper bounds on the eigenvalues of the matrix $A(c)$ given by (1.1). If an optimization method based on active sets of inequality constraints is used, i.e., where the inequalities thought to be binding are replaced by equality constraints, one has a problem closely related to Problem 1. (The same remark applies to the educational testing problem, which has constraints of the form $\lambda_i(x) \geq 0$.) It is interesting to note that multiple eigenvalues will naturally tend to occur at a solution, since the minimization objective may drive several eigenvalues to the same bound. It is therefore very important to handle multiple eigenvalues correctly. We will explain how to do this in §3, in the context of Problem 1. A closely related problem has the simple form

$$\begin{aligned} &\min u \\ &\text{subject to} \quad \lambda_i(c) \leq u \quad i = 1, \ldots, n, \end{aligned} \tag{1.10}$$

i.e. the object is to minimize $\lambda_n(c)$, the maximum eigenvalue of $A(c)$. Again, multiple eigenvalues generally occur at the solution. There is a large engineering literature on problems of the form (1.9) and (1.10). See, for example, Polak and Wardi (1982) and Mayne and Polak(1982) for a discussion of problems from control theory involving restrictions on singular values of a transfer matrix, and Olhoff and Taylor (1983) for a discussion of problems from structural analysis and optimal design.

Another interesting variation is the *graph partitioning problem*, given by Cullum, Donath and Wolfe (1975). The objective is to minimize the sum of the largest eigenvalues of a symmetric matrix, as a function of its diagonal entries. More precisely, consider the problem

$$\min_D \varphi(D) = \sum_{i=n-r}^{n} \lambda_i(A_0 + D)$$

$$\text{subject to} \quad trace(D) = 0,$$

where the symmetric matrix A_0 and the integer r are given, and D is diagonal. This problem can be transformed into

$$\min_{D,u} \sum_{i=n-r}^{n} u_i$$

$$\text{subject to} \quad trace(D) = 0$$

$$\lambda_i(A_0 + D) \leq u_i \qquad n - r \leq i \leq n.$$

One therefore has a problem closely related to (1.9).

There are several inverse eigenvalue problems with special structure that can be solved by direct methods. An example of this is the reconstruction of Jacobi matrices from spectral data; see de Boor and Golub(1978). We will not consider these types of problems.

§1.2 Notation and Definitions.

We define $\lambda(c) = [\lambda_1(c), \ldots, \lambda_n(c)]^T$ and $\Lambda(c) = diag(\lambda_i(c))$. A solution to Problem 1 will be denoted by c^*, and we write $\lambda^* = [\lambda_1^*, \ldots, \lambda_n^*]^T$ and $\Lambda^* = diag(\lambda_i^*)$. Since $A(c)$ is symmetric, it has an orthonormal set of eigenvectors $\{q_i(c)\}_1^n$. The orthogonal matrix $Q(c) = [q_1(c), \ldots, q_n(c)]$ will be called a *matrix of eigenvectors* of $A(c)$. Throughout the paper $\|\cdot\|$ denotes the Euclidean vector norm ($\ell_2 - norm$) or its corresponding induced matrix norm, and $\|\cdot\|_F$ the Frobenius matrix norm.

§2. DISTINCT EIGENVALUES

We will now describe several methods for solving Problem 1 in the case where the given eigenvalues are distinct. In §3 we will see how to cope with multiple eigenvalues. Assume there exists a solution c^* to Problem 1. Then there is a neighborhood of c^* where the eigenvalues $\lambda_i(c)$ are distinct and are differentiable functions (see for example Ortega (1972), p.54). In this neighborhood we will consider the nonlinear system of equations

$$f(c) = \begin{bmatrix} \lambda_1(c) - \lambda_1^* \\ \vdots \\ \lambda_n(c) - \lambda_n^* \end{bmatrix} = 0. \tag{2.1}$$

The first method we will describe consists of applying Newton's method to (2.1). Differentiating the relations

$$q_i(c)^T q_i(c) = 1 \tag{2.2}$$

$$q_i(c)^T A(c) q_i(c) = \lambda_i(c), \tag{2.3}$$

one finds that

$$\frac{\partial \lambda_i(c)}{\partial c_k} = q_i(c)^T A_k q_i(c). \tag{2.4}$$

Thus the Jacobian of f is

$$J_{ik}(c) = q_i(c)^T A_k q_i(c), \tag{2.5}$$

and one step of Newton's method is defined by

$$J(c^\nu)(c^{\nu+1} - c^\nu) = -f(c^\nu). \tag{2.6}$$

We will write (2.6) in a slightly different form. From (1.1) and (2.3) we have

$$q_i(c)^T A_0 q_i(c) + \sum_{k=1}^{n} q_i(c)^T A_k q_i(c) c_k = \lambda_i(c),$$

and therefore (2.6) can be written as

$$q_i(c^\nu)^T A(c^{\nu+1}) q_i(c^\nu) = \lambda_i^*, \tag{2.7}$$

or equivalently,

$$J(c^\nu)c^{\nu+1} = \lambda^* - b(c^\nu) \tag{2.8}$$

where

$$b_i(c) = q_i(c)^T A_0 q_i(c) \qquad i = 1, \ldots, n. \tag{2.9}$$

Thus Newton's method for solving (2.1) is:

Method I.

Choose a starting value c^0. Form $A(c^0)$ and find its eigenvalues and eigenvectors.

For $\nu = 0, 1, 2, \ldots$

1. Stop if $\|\lambda(c^\nu) - \lambda^*\|$ is sufficiently small.
2. Form $J(c^\nu)$ (see (2.5)) and $b(c^\nu)$ (see (2.9)) and compute $c^{\nu+1}$ by solving (2.8).Form $A(c^{\nu+1})$.
3. Find the eigenvalues $\{\lambda_i(c^{\nu+1})\}$ and eigenvectors $\{q_i(c^{\nu+1})\}$ of $A(c^{\nu+1})$.

Method I has been studied by many authors. An early reference is Downing and Householder (1956), where the method is proposed for solving the additive inverse and multiplicative inverse eigenvalue problems. Physicists have used it for many years in nuclear spectroscopy calculations (see Brussard and Glaudemans (1977)). Kublanovskaja (1970) has given a convergence analysis of this method.

Describing the iteration by means of (2.6) seems more natural than using (2.8). However, the latter has the same form as the next two methods we will present below. Also (2.8) shows that the direction produced by Newton's method does not depend explicitly on the eigenvalues $\lambda(c^\nu)$.

Instead of computing the eigenvectors of $A(c)$ at each step we may consider approximating them. One possibility is to use inverse iteration. Suppose that c^ν is our current estimate of the parameters and $Q^{(\nu)}$ is an approximation to $Q(c^\nu)$, the matrix of eigenvectors of $A(c^\nu)$. Let q_i^ν be the i^{th} column of $Q^{(\nu)}$. To compute a new estimate $c^{\nu+1}$ we form

$$J_{ik}^{(\nu)} = (q_i^\nu)^T A_k q_i^\nu \qquad i, k = 1, \ldots, n \tag{2.10}$$

$$b_i^\nu = (q_i^\nu)^T A_0 q_i^\nu \qquad i = 1, \ldots, n \tag{2.11}$$

and solve

$$J^{(\nu)}c^{\nu+1} = \lambda^* - b^\nu. \tag{2.12}$$

(Compare with (2.5), (2.9) and (2.8)). To update our approximations to the eigenvectors we apply one step of inverse iteration: we compute $\gamma_i, i = 1, \ldots, n$ by solving

$$[A(c^{\nu+1}) - \lambda_i^* I]\gamma_i = q_i^\nu \qquad i = 1, \ldots, n. \tag{2.13}$$

We then define

$$q_i^{\nu+1} = \frac{\gamma_i}{\|\gamma_i\|} \qquad i = 1, \ldots, n,$$

which determine the new matrix $Q^{(\nu+1)}$. Thus we are performing a Newton-like iteration where instead of computing the exact eigenvectors of $A(c)$ at each step we update an approximation to them by performing one step of inverse iteration.

Method II.

Choose a starting value c^0. Form $A(c^0)$ and compute its matrix of eigenvectors $Q(c^0)$. Set $Q^{(0)} \leftarrow Q(c^0)$.

For $\nu = 0, 1, 2, \ldots$

1. If $\|(Q^{(\nu)})^T A(c^\nu)Q^{(\nu)} - \Lambda^*\|_F$ is sufficiently small, stop.
2. Form $J^{(\nu)}$ (see (2.10)) and b^ν (see (2.11)) and compute $c^{\nu+1}$ by solving (2.12). Form $A(c^{\nu+1})$.
3. Compute the factorization

$$A(c^{\nu+1}) = UNU^T,$$

where U is orthogonal and N is tridiagonal. Solve the n linear systems

$$[N - \lambda_i^* I](U^T\gamma_i) = U^T q_i^\nu \qquad i = 1, \ldots, n$$

and compute

$$q_i^{\nu+1} = \frac{\gamma_i}{\|\gamma_i\|}.$$

Method II is closely related to a method proposed by Osborne (1971); see also Hald (1972). We have used a different right-hand side vector in (2.12) to take advantage of the fact that $A(c)$ is an affine function. As we will discuss below, the form of the method proposed by Osborne can be useful when $A(c)$ is a nonlinear function.

A different approach is based on the use of matrix exponentials and Cayley transforms. A solution to Problem 1 can be described by c and Q, where Q is an orthogonal matrix and

$$Q^T A(c) Q = \Lambda^*. \tag{2.14}$$

Suppose that $Q^{(\nu)}$ is our current estimate of Q. Let us write $Q = Q^{(\nu)} e^Y$ where Y is a skew-symmetric matrix, i.e., $Y^T = -Y$. Then (2.14) can be written as

$$\begin{aligned}(Q^{(\nu)})^T A(c) Q^{(\nu)} &= e^Y \Lambda^* e^{-Y} \\ &= (I + Y + \tfrac{1}{2}Y^2 + \ldots)\Lambda^*(I - Y + \tfrac{1}{2}Y^2 + \ldots).\end{aligned}$$

Thus

$$(Q^{(\nu)})^T A(c) Q^{(\nu)} = \Lambda^* + Y\Lambda^* - \Lambda^* Y + O(\|Y\|^2). \tag{2.15}$$

We now define a new estimate of the parameters $c^{\nu+1}$ by neglecting second-order terms in Y and equating the diagonal elements of (2.15). We obtain

$$(q_i^\nu)^T A(c^{\nu+1}) q_i^\nu = \lambda_i^* \qquad i = 1\ldots, n, \tag{2.16}$$

which is identical to (2.12) with $J^{(\nu)}$ and $b^{(\nu)}$ defined by (2.10) and (2.11). Thus the new estimate of $c^{\nu+1}$ is obtained in the same way as in Methods I and II. Equating the off-diagonal elements of (2.15), with second order terms neglected, we have

$$y_{ij}(\lambda_j^* - \lambda_i^*) = (q_i^\nu)^T A(c^{\nu+1}) q_j^\nu \qquad 1 \le i < j \le n. \tag{2.17}$$

The matrix Y is completely determined by (2.17), since $Y = -Y^T$ and we are assuming that $\{\lambda_i^*\}$ are distinct. Now construct an orthogonal matrix P using the Cayley transform

$$P = (I + \tfrac{1}{2}Y)(I - \tfrac{1}{2}Y)^{-1} \tag{2.18}$$

and compute the new estimate of the matrix of eigenvectors by

$$Q^{(\nu+1)} = Q^{(\nu)}P. \tag{2.19}$$

As we neglected second order terms, $Q^{(\nu+1)}$ is only an approximation to the desired matrix and we need to iterate.

Method III.

Choose a starting value c^0. Form $A(c^0)$ and compute its matrix of eigenvectors $Q(c^0)$. Set $Q^{(0)} \leftarrow Q(c^0)$.

For $\nu = 0, 1, 2, \ldots$

1. If $\|(Q^{(\nu)})^T A(c^\nu)Q^{(\nu)} - \Lambda^*\|_F$ is sufficiently small, stop.
2. Form $J^{(\nu)}$ (see (2.10)) and b^ν (see (2.11)) and compute $c^{\nu+1}$ by solving (2.12). Form $A(c^{\nu+1})$.
3. For $i = 1, \ldots, n$ and $j = i+1, \ldots, n$, compute

$$y_{ij} = \frac{(q_i^\nu)^T A(c^{\nu+1}) q_j^\nu}{\lambda_j^* - \lambda_i^*}.$$

4. (Note that (2.18) and (2.19) imply

$$(Q^{(\nu+1)})^T = (I + \tfrac{1}{2}Y)^{-1}(I - \tfrac{1}{2}Y)(Q^{(\nu)})^T.)$$

Compute $H = (I - \frac{1}{2}Y)(Q^{(\nu)})^T)$. Let h_i be the i^{th} column of H.

Factorize the matrix $I + \frac{1}{2}Y$, and use this to solve the n linear systems

$$(I + \tfrac{1}{2}Y)v_i = h_i \qquad i = 1, \ldots, n.$$

Set

$$(Q^{(\nu+1)})^T = [v_1, \ldots, v_n].$$

This method is apparently new. Downing and Householder(1956) use the Cayley transform to motivate the Newton step (2.6) in Method I. However, they do not suggest updating approximations to the eigenvectors, but instead compute the exact eigenvectors of $A(c)$ at each step.

One can motivate Method III following a different reasoning. Suppose that we are given an initial matrix $B^{(0)}$ whose eigenvalues coincide with the target eigenvalues $\{\lambda_i^*\}$. If $B^{(0)}$ can be written in the form (1.1), the problem is solved. Otherwise we generate a sequence $\{B^{(\nu)}\}, \nu = 1, 2, \ldots$, which converges to a matrix of the form (1.1), and with the property that each matrix in the sequence has a spectrum that coincides with the target spectrum. Given $B^{(\nu)}$, we would like to find a skew-symmetric matrix Z and a vector $c^{\nu+1}$ such that

$$e^{-Z} B^{(\nu)} e^{Z} = A(c^{\nu+1}).$$

Expanding e^Z and neglecting second order terms we obtain

$$B^{(\nu)} + B^{(\nu)} Z - Z B^{(\nu)} = A(c^{\nu+1}).$$

The diagonal equations determine $c^{(\nu+1)}$, as before, and the off-diagonal equations determine Z. We now let $R = (I + \frac{1}{2}Z)(I - \frac{1}{2}Z)^{-1}$ and define $B^{(\nu+1)} = R^T B^{(\nu)} R$. To find $B^{(0)}$ we may proceed as follows. Let c^0 be our first estimate of the parameters. Compute the eigenvectors $\{q_i(c_0)\}$ of $A(c^0)$ and define

$$B^{(0)} = \sum_{i=1}^{n} \lambda_i^* q_i(c^0) q_i(c^0)^T.$$

Then $B^{(0)}$ has the target spectrum and its eigenvectors coincide with those of $A(c^0)$. It is not difficult to see that this process is identical to Method III.

Let us now look at a different formulation of Problem 1. Consider the nonlinear system

$$g(c) = \begin{bmatrix} det\big(A(c) - \lambda_1^* I\big) \\ \vdots \\ det\big(A(c) - \lambda_n^* I\big) \end{bmatrix} = 0. \tag{2.20}$$

Note that the i^{th} equation of (2.20) can be written as

$$g_i(c) = \prod_{k=1}^{n} \left(\lambda_k(c) - \lambda_i^*\right). \tag{2.21}$$

To apply Newton's method to this new system we first need to compute the Jacobian. From (2.4) and (2.21) it follows that

$$\frac{\partial g_i(c)}{\partial c_j} = \sum_{k=1}^{n} \left[q_k^T A_j q_k \prod_{\substack{\ell=1 \\ \ell \neq k}}^{n} \left(\lambda_\ell(c) - \lambda_i^*\right)\right]. \tag{2.22}$$

Therefore the Jacobian of g is

$$G(c) = \begin{bmatrix} \prod\limits_{\ell \neq 1} \left(\lambda_\ell(c) - \lambda_1^*\right) & \cdots & \prod\limits_{\ell \neq n} \left(\lambda_\ell(c) - \lambda_1^*\right) \\ \vdots & & \vdots \\ \prod\limits_{\ell \neq 1} \left(\lambda_\ell(c) - \lambda_n^*\right. & \cdots & \prod\limits_{\ell \neq n} \left(\lambda_\ell(c) - \lambda_n^*\right) \end{bmatrix} \times$$

$$\begin{bmatrix} q_1(c)^T A_1 q_1(c) & \cdots & q_1(c)^T A_n q_1(c) \\ \vdots & & \vdots \\ q_n(c)^T A_1 q_n(c) & \cdots & q_n(c)^T A_n q_n(c) \end{bmatrix}$$

$$= diag\left(g_i(c)\right) \begin{bmatrix} \dfrac{1}{\lambda_1(c) - \lambda_1^*} & \cdots & \dfrac{1}{\lambda_n(c) - \lambda_1^*} \\ \vdots & & \vdots \\ \dfrac{1}{\lambda_1(c) - \lambda_n^*} & \cdots & \dfrac{1}{\lambda_n(c) - \lambda_n^*} \end{bmatrix} J(c)$$

$$= diag\left(g_i(c)\right) \cdot diag\left(\frac{1}{f_i(c)}\right) \cdot V(c) \tag{2.23}$$

where $J(c)$ is given by (2.5) and $V(c)$ is defined by

$$V_{ij}(c) = \sum_{k=1}^{n} \frac{\left[q_k(c)^T A_j q_k(c)\right]\left[\lambda_i(c) - \lambda_i^*\right]}{\lambda_k(c) - \lambda_i^*}. \tag{2.24}$$

The Newton iterate is therefore

$$c^{\nu+1} = c^\nu - V(c^\nu)^{-1} diag\left(f_i(c^\nu)\right) diag\left(\frac{1}{g_i(c^\nu)}\right) g(c^\nu)$$

$$= c^\nu - V(c^\nu)^{-1} f(c^\nu).$$

Method IV.

Choose a starting value c^0. Form $A(c^0)$ and compute its eigenvalues and eigenvectors.

For $\nu = 0, 1, \ldots$

1. Stop if $\|\lambda(c^\nu) - \lambda^*\|$ is sufficiently small.
2. Form $V(c^\nu)$ (see (2.24)) and compute $c^{\nu+1}$ by solving

$$V(c^\nu)(c^{\nu+1} - c^\nu) = -f(c^\nu).$$

Form $A(c^{\nu+1})$.

3. Find the eigenvalues $\{\lambda_i(c^{\nu+1})\}_1^n$ and eigenvectors $\{q_i(c^{\nu+1})\}_1^n$ of $A(c^{\nu+1})$.

This method was proposed by Biegler-Kōnig (1981) and generalizes an algorithm of Lancaster (1964-a). To show its relation to Method I we note that $V(c)$ can be written as

$$V(c) = W(c) \cdot J(c),$$

where the matrix W is defined by

$$W(c) = \begin{bmatrix} 1 & \dfrac{\lambda_1(c) - \lambda_1^*}{\lambda_2(c) - \lambda_1^*} & \cdots & \dfrac{\lambda_1(c) - \lambda_1^*}{\lambda_n(c) - \lambda_1^*} \\ \vdots & \vdots & & \vdots \\ \dfrac{\lambda_n(c) - \lambda_n^*}{\lambda_1(c) - \lambda_n^*} & \dfrac{\lambda_n(c) - \lambda_n^*}{\lambda_2(c) - \lambda_n^*} & \cdots & 1 \end{bmatrix}. \tag{2.25}$$

Thus Method IV differs from Method I in that $J(c^\nu)$ has been replaced by $W(c^\nu)J(c^\nu)$. Since the given eigenvalues $\{\lambda_i^*\}_1^n$ are distinct, $W(c) \to I$ as $c \to c^*$, and so asymptotically Methods I and IV coincide. Nevertheless, our numerical experience indicates that Method I almost always requires fewer iterations, and that Method IV suffers more often from ill-conditioning. One can readily see the drawback of using formulation (2.20) by noting that

$$g_i(c) = f_i(c) \prod_{k \neq i} (\lambda_k(c) - \lambda_i^*). \tag{2.26}$$

One is thus complicating system (2.1) by multiplying each equation by a polynomial of degree $n - 1$. Suppose, for example, that the problem is so simple that the functions

$\{\lambda_i(c)\}_1^n$ are linear. Then (2.1) is a linear system of equations and Method I will find the solution in one step. On the other hand, (2.20) represents a system of polynomial equations of degree n, and Method IV will have to iterate to approach the solution. It therefore seems that Method I is always to be preferred over Method IV.

Let us now compare the computational requirements of Methods I,II and III. We will first assume that $A(c)$ is dense. The computation of the eigenvalues and eigenvectors in Method I requires approximately $5n^3$ multiplication operations; see Golub and Van Loan (1983). Method II requires approximately $3n^3$ multiplications to update the q_i , whereas Method III requires approximately $4n^3$ multiplications in Steps 3 and 4. Note that all the methods require n^4 multiplications to form the matrix J in Step 2. However, in the case of the additive inverse problem (1.2), forming J requires only n^2 operations. If the matrix $A(c)$ is sparse, Method III becomes less competitive. For example, if $A(c)$ is tridiagonal, Method I requires only about $14n^2$ multiplications per iteration ($9n^2$ to compute the eigenvalues (Parlett (1980, p.165)) plus $5n^2$ to find the eigenvectors by inverse iteration), while Method II requires $5n^2$ multiplications in Step 3 and Method III requires about $3n^3$ multiplications in Steps 3 and 4. In §4 we will comment on the numerical behavior of the three methods.

Methods I, II, III and IV are locally quadratically convergent under some nonsingularity assumptions. This will be shown in §3.3. There are, on the other hand, various methods for solving Problem 1 that are only linearly convergent. One of these methods was proposed by Downing and Householder (1956). The iteration is

$$c^{\nu+1} = c^{\nu} - \big(\lambda(c) - \lambda^*\big),$$

and thus is obtained from Method I by replacing $J(c^{\nu})$ by the identity matrix. A different method, specifically designed for solving the additive inverse eigenvalue problem (1.2), was proposed by Hald (1972). Suppose that $D^{(\nu)}$ is our current estimate of the desired diagonal matrix, and that $Q^{(\nu)}$ is the matrix of eigenvectors of $A_0 + D^{(\nu)}$. We define $D^{(\nu+1)}$ as the diagonal matrix that solves the problem

$$\min_D \|(A_0 + D)Q^{(\nu)} - Q^{(\nu)}\Lambda^*\|_F.$$

Since

$$\|(A_0 + D)Q^{(\nu)} - Q^{(\nu)}\Lambda^*\|_F = \|D - (Q^{(\nu)}\Lambda^*(Q^{(\nu)})^T - A_0)\|_F$$

it is clear that

$$D^{(\nu+1)} = diag\big((Q^{(\nu)}\Lambda^*(Q^{(\nu)})^T - A_0)_{ii}\big). \tag{2.27}$$

Thus in this method one computes the eigenvectors at each step and updates the estimate of D by means of (2.27). Friedland (1977) generalized this method for more general functions $A(c)$.

A method closely related to (2.27) can be used for solving the problem of communality. Recall that the problem is: given A_0, find a diagonal matrix D such that $A_0 + D$ has as small rank as possible. We start by making a guess of the rank: say that it is $n - t$. Thus t eigenvalues will be zero at the solution. Suppose that $D^{(\nu)}$ is our current estimate, and let $Q^{(\nu)}$ and $\Lambda^{(\nu)}$ be the matrices of eigenvectors and eigenvalues of $A_0 + D^{(\nu)}$. We define

$$D^{(\nu+1)} = diag\big((Q^{(\nu)}\overline{\Lambda}^{(\nu)}(Q^{(\nu)})^T - A_0)_{ii}\big)$$

where $\overline{\Lambda}^{(\nu)}$ is obtained from $\Lambda^{(\nu)}$ by setting the t smallest diagonal elements (in absolute value) to zero. Ideally we would like to use Λ^* instead of $\overline{\Lambda}^{(\nu)}$, but only t zero elements of Λ^* are known. This method is described in Holzinger and Harman (1941), and our numerical experience indicates that it is robust but very slowly convergent.

We will now discuss what changes are needed in the numerical methods described so far, when $A(c)$ is not affine, but is a nonlinear function of c. Note that (2.4) should be replaced by

$$\frac{\partial \lambda_i(c)}{\partial c_k} = q_i(c)^T A_k(c) q_i(c) \tag{2.28}$$

where

$$A_k(c) = \frac{\partial}{\partial c_k} A(c). \tag{2.29}$$

Method I is then defined by (2.6) where $J(c^\nu)$ is formed by using (2.28). Note that (2.8) cannot be used since it was derived under the assumption that $A(c)$ is affine. Similarly, for Method IV we need only replace A_j by $A_j(c)$ in (2.24).

For Method II we do not wish to compute the eigenvalues $\lambda_i(c)$ required for (2.6). A natural modification to Method II is to consider an iteration of the form (2.6), where the vectors q_i are updated by using the inverse iteration (2.13), and the eigenvalues $\lambda_i(c)$ are approximated by means of the Rayleigh quotient, i.e., $\lambda_i(c) \cong q_i^T A(c) q_i$. A different approach was suggested by Osborne (1971) and does not require approximating the eigenvalues $\lambda_i(c)$ explicitly. He defines the function $\beta(c) = [\beta_1(c), \ldots, \beta_n(c)]^T$ by

$$\beta_i(c) = (\gamma_i^T q_i)^{-1} \qquad i = 1, \ldots, n,$$

where $\{q_i\}_1^n$ are our current approximations to the eigenvectors and the γ_i are given by

$$[A(c) - \lambda_i^* I]\gamma_i = q_i \qquad i = 1, \ldots, n.$$

If we apply one step of Newton's method to the system $\beta(c) = 0$ we obtain an iteration of the form

$$\widetilde{J}(c^\nu)(c^{\nu+1} - c^\nu) = -\beta(c^\nu),$$

where

$$\widetilde{J}_{ik}(c^\nu) = \left((\gamma_i^\nu)^T A_k(c^\nu)\gamma_i^\nu\right)\beta_i(c^\nu)^2$$

and $A_k(c)$ is defined by (2.29). It is not difficult to see that $\beta(c)$ approximates $\lambda(c) - \lambda^*$; see Hald (1972) for details.

Finally, consider Method III. Equation (2.16) is now a nonlinear equation in c. We can, however, replace $A(c^{\nu+1})$ in (2.16) by the first order approximation

$$A(c^\nu) + A_1(c^\nu)(c_1^{\nu+1} - c_1^\nu) + \cdots + A_n(c^\nu)(c_n^{\nu+1} - c_n^\nu)$$

to obtain

$$\widehat{J}(c^\nu)(c^{\nu+1} - c^\nu) = -(\widehat{\lambda}^\nu - \lambda^*),$$

where

$$\widehat{J}_{ik}(c^\nu) = (q_i^\nu)^T A_k(c) q_i^\nu$$

and $\widehat{\lambda}^\nu = [\widehat{\lambda}_1^\nu, \ldots, \widehat{\lambda}_n^\nu]^T$ is defined by

$$\widehat{\lambda}_i^\nu = (q_i^\nu)^T A(c^\nu) q_i^\nu,$$

the Rayleigh quotient. After $c^{\nu+1}$ has been computed we update the vectors q_i by (2.17)–(2.19).

§3. MULTIPLE EIGENVALUES

In this section we suppose that $\{\lambda_i^*\}$ includes multiple eigenvalues, and that a solution c^* to Problem 1 exists. We first describe how the methods of §2, without modifications, behave in this case, and then explain how the problem formulation and methods should be changed when multiple eigenvalues are present. For convenience we assume that only the first eigenvalue is multiple, with multiplicity t, *i.e.*,

$$\lambda_1^* = \lambda_2^* = \cdots = \lambda_t^* < \lambda_{t+1}^* < \cdots < \lambda_n^*.$$

There is no difficulty in generalizing all our remarks to an arbitrary set of given eigenvalues.

§3.1 Behavior of Unmodified Methods.

Let us first consider Method I. When the given eigenvalues are distinct it is straightforward to see that Method I is locally quadratically convergent, under the assumption that the Jacobian (2.5) is nonsingular at c^*. The reason is that Method I consists of applying

Newton's method for finding a zero of f, defined by (2.1), which is a smooth function. In fact, the first partial derivatives of $\lambda_i(c)$ are given by (2.4), and it is not difficult to show that

$$\frac{\partial^2 \lambda_i}{\partial c_k \partial c_j} = 2 \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} \frac{[q_\ell(c)^T A_k q_i(c)][q_\ell(c)^T A_j q_i(c)]}{\lambda_i(c) - \lambda_\ell(c)} \tag{3.1}$$

(see for example Lancaster (1964-b)). Thus f satisfies the well-known conditions for quadratic convergence of Newton's method: (i) f is differentiable and $J(c)$ is Lipschitz continuous in a neighborhood of c^*; (ii) $J(c^*)$ is nonsingular. (See Ortega and Rheinboldt (1970), p. 312).

However, we can see from (3.1) that as the separation of the eigenvalues decreases, the Lipschitz constant generally grows, the problem becomes increasingly ill-conditioned, and the neighborhood of the solution in which convergence takes place becomes smaller. When the separation is zero, *i.e.*, when multiple eigenvalues are present, the eigenvalues are not, in general, differentiable at c^*. Furthermore, the eigenvectors $\{q_i(c^*)\}$ are not unique, and they cannot generally be defined to be continuous functions of c at c^*. This can be seen by considering the example $A_0 = 0$,

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad c^* = \lambda^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Note that as long as the eigenvalues of $A(c^\nu)$ are distinct for all iterates c^ν, Method I remains well-defined. However, the matrix of eigenvectors $Q(c^\nu)$, and consequently the Jacobian $J(c^\nu)$, generally will not converge as $c^\nu \to c^*$. Therefore one might expect that in the multiple eigenvalue case the method is at best slowly convergent to c^*. In fact, however, the convergence is generally quadratic, both in theory and in practice. This fact was apparently first established by Nocedal and Overton (1983), although Bohte (1967-68) had observed earlier that the method experienced no difficulties in his numerical tests with multiple eigenvalues. The quadratic convergence may be explained in several ways. Nocedal and Overton (1983) base their analysis on a classical result of Rellich (1969), which states that the eigenvalues can be defined to be analytic functions of a single variable, along

any line passing through the solution c^*. By using the mean value theorem in one variable it follows that, locally, every Newton step produces a quadratic contraction in the error. The result is that, given a nonsingularity assumption, the iterates $\{c^\nu\}$ converge quadratically although the sequence $\{J(c^\nu)\}$ does not converge. A completely different proof of this result will be given in §3.3.

We comment further here on the nonsingularity assumption needed for quadratic convergence. It is sufficient to assume that $\{A(c^\nu)\}$ has distinct eigenvalues for all ν and that $\{\|J(c^\nu)^{-1}\|\}$ is bounded. Even though this condition usually holds in practice, it would be more desirable to state the nonsingularity assumption in terms of a matrix evaluated at c^*. Since the matrix of eigenvectors Q is not uniquely determined at c^*, let us define

$$\Omega = \{Q : Q^T Q = I \text{ and } Q^T A(c^*) Q = \Lambda^*\}$$

and, for any $Q \in \Omega$, define $J^*(Q)$ by

$$J^*_{ik}(Q) = q_i^T A_k q_i.$$

To obtain a useful nonsingularity assumption we may consider

$$\sup_{Q \in \Omega} \{\|J^*(Q)^{-1}\|\}.$$

However, it turns out that, in general, this supremum does not exist. By solving a system of $n - t + 1$ linear equations in n unknowns and doing an appropriate transformation, it is generally possible to choose the eigenvectors so that $J^*(Q)$ is singular. To see this, let $Q = [q_1, \ldots, q_n]$ be any set of eigenvectors of $A(c^*)$. Consider the system

$$\sum_{k=1}^{n} \left(\sum_{i=1}^{t} q_i^T A_k q_i \right) x_k = 0$$

$$\sum_{k=1}^{n} \left(q_i^T A_k q_i \right) x_k = 0 \qquad i = t+1, \ldots, n,$$

where the unknown $x = [x_1, \ldots, x_n]^T \in \mathrm{R}^n$. This homogeneous system is always solvable for some $x \neq 0$, assuming $t > 1$. Let $Q_1 = [q_1, \ldots, q_t]$. By construction,

$$tr\big(Q_1^T(A(x) - A_0)Q_1\big) = 0 = q_i^T\big(A(x) - A_0\big)q_i, \qquad i = t+1, \ldots, n.$$

It follows that $Q_1^T\big(A(x) - A_0\big)Q_1$ is orthogonally similar to a matrix with zero diagonal (this is readily proven by induction on n). Let U define this similarity transformation, let $\widehat{Q}_1 = Q_1 U$, and define $\widehat{Q} = [\widehat{Q}_1, q_{t+1}, \ldots, q_n]$. Then, by construction, $J^*(\widehat{Q})x = 0$ which implies that $J^*(\widehat{Q})$ is singular.

The question then arises: does there exist a direction d such that $A(c^* + \alpha d)$ has distinct eigenvalues for small $\alpha > 0$ and such that the condition

$$Q\big(A(c^* + \alpha d)\big) \to \widehat{Q} \quad \text{as} \quad \alpha \to 0^+ \tag{3.2}$$

holds? The answer is, generically, yes, provided $n \geq t(t-1)/2$. To obtain such a vector d, one solves

$$\sum_{k=1}^{n} (\widehat{Q}_1^T A_k \widehat{Q}_1) d_k - diag(\mu_i) = 0,$$

a system of $t(t+1)/2$ equations in $n + t$ unknowns d and $\{\mu_i\}_1^t$. When $n \geq t(t-1)/2$, there is generally a solution with $\{\mu_i\}_1^t$ distinct. In this case it is not difficult to show that (3.2) holds.

The consequence of the discussion above is that generally there is a manifold $\mathcal{M}$ of dimension less than n for which $J(c)$ approaches singularity when $c \in \mathcal{M}$ and $c \to c^*$. By carrying out the construction described above, one is able to obtain a point $c \in \mathcal{M}$, with c near c^*. Interestingly enough, however, even when Method I is started at such a point, it has no difficulty with convergence. Typically the next iterate c^1 is not in or near $\mathcal{M}$, and subsequent iterates converge quadratically as usual. Because $\mathcal{M}$ has dimension less than n, it is very unlikely that iterates in $\mathcal{M}$ will be generated. One might well be concerned at the possiblity that if iterates near $\mathcal{M}$ are generated, convergence could be slow. However, one must remember that $J(c)$ is not continuous at c^*, and that, for example, $J(c)$ varies

extremely rapidly on small spheres centered at c^*. Thus even at a point c lying near $\mathcal{M}$, and near c^*, $J(c)$ may be far from singular.

So far we have explained the behavior of Method I in the presence of multiple eigenvalues without making any modification to the method. In §3.2 we shall see that when the problem is properly formulated, the method should be modified. In that case J is replaced by a matrix which does not have the undesirable property that its nonsingularity depends on the choice of the eigenvectors. The interesting conclusion from the discussion just completed, however, is that even when no modifications are made to Method I, it is generally locally quadratically convergent regardless of the eigenvalue multiplicities, assuming the problem has a solution.

Let us now consider Method II. Difficulties may arise during the application of the inverse iteration (2.13) due to the presence of the multiple eigenvalue λ_1^*. To see this, let us write

$$q_i^\nu = \sum_{k=1}^{n} \alpha_k^{(i)} q_k(c^{\nu+1}) \qquad i = 1, \ldots, n,$$

and thus (2.13) gives

$$\gamma_i = \sum_{k=1}^{n} \frac{\alpha_k^{(i)}}{\lambda_k(c^\nu) - \lambda_i^*} q_k(c^{\nu+1}) \qquad i = 1, \ldots, n.$$

Now suppose, for example, that λ_1^* is much closer to $\lambda_1(c^\nu)$ than it is to $\lambda_2(c^\nu), \ldots, \lambda_t(c^\nu)$. Then all the vectors $\{\gamma_i\}_1^t$ will be nearly parallel to $q_1(c^{\nu+1})$ and will fail to approximate the invariant subspace. To avoid this difficulty each new vector γ_i, $1 \le i \le t$, is orthogonalized with respect to the earlier computed vectors belonging to the multiple eigenvalue. Thus we solve

$$[A(c^{\nu+1}) - \lambda_1^* I]\, \Gamma = Q_1^{(\nu)}$$

for $\Gamma \in \mathbf{R}^{n \times t}$, where $Q_1^{(\nu)} = [q_1^\nu, \ldots, q_t^\nu]$, and then compute the "$QR$" factorization

$$\Gamma = Q_1^{(\nu+1)} T,$$

where T is upper triangular. If the orthogonalization produces a zero vector, we replace the corresponding column of $Q_1^{(\nu)}$ by a unit vector e_j, trying the columns of the identity matrix in turn until we succeed (see Peters and Wilkinson (1971) p.418). The vectors corresponding to the distinct eigenvalues are updated by means of (2.13). It will be shown in §3.3 that Method II, using this implementation of the inverse iteration, is locally quadratically convergent. The same argument given for Method I shows that $\{\|J^{(\nu)-1}\|\}$ may not be bounded. However, as in Method I, this is very unlikely to occur; furthermore, this consideration will disappear when we modify the methods in §3.2.

We now turn our attention to Method III. The diagonal equations (2.16) define $c^{\nu+1}$ just as Methods I and II do, regardless of the eigenvalue multiplicities. The off-diagonal equations (2.17) are

$$y_{ij}(\lambda_j^* - \lambda_i^*) = (q_i^\nu)^T A(c^{\nu+1})q_i^\nu \qquad 1 \le i < j \le n.$$

The left-hand side of this equation is zero for $1 \le i < j \le t$ regardless of the value of y_{ij}, and thus y_{ij} is not determined. A reasonable course to take is to set

$$y_{ij} = 0 \qquad 1 \le i < j \le t.$$

With this choice it will be shown in §3.3 that the iterates $\{c^\nu\}$ converge quadratically to c^* as in Methods I and II. Furthermore, it will be shown that $\{Q^{(\nu)}\}$ converges quadratically to a limit. It follows that $\{J^{(\nu)}\}$ converges. The remarks made for Methods I and II, concerning the possible unboundedness of $\{\|J^{(\nu)-1}\|\}$, also apply to Method III.

Now consider Method IV. Here the analysis is trivial. In the multiple eigenvalue case, the function $g(c)$ remains differentiable, but has multiple entries. The Jacobian of g is thus necessarily singular. It is clear that the method must be reformulated. However, since we do not consider this method computationally attractive, we will not discuss it any further.

§3.2 Modification to Formulation and Methods.

Let us view the problem from a slightly different perspective. The relation

$$A(c) = Q\Lambda^* Q^T \tag{3.3}$$

can be considered a system of $n(n+1)/2$ equations in $n(n+1)/2$ unknowns, namely the parameters $\{c_k\}_1^n$ and the orthogonal matrix Q which has $n(n-1)/2$ degrees of freedom. However, when $t > 1$, $s = t(t-1)/2$ of these degrees of freedom are of no help in solving the problem (3.3), since they describe only the rotation of $[q_1, \ldots, q_t]$ to a different choice of basis. It follows that when Λ^* is completely specified and $t > 1$, Problem 1 is inherently overdetermined. This did not cause difficulties for our discussion in §3.1 because we assumed throughout that a solution c^* existed, *i.e.*, that Λ^* was choosen so that (3.3) was solvable. However, in practice we must expect that if a multiple eigenvalue is present, either s of the remaining eigenvalues are not specified, or an additional s parameters are available. For convenience we shall make the former assumption and, instead of Problem 1, consider

PROBLEM 2. *Find the parameters $c_1, \ldots, c_n$ so that the $n - s$ smallest eigenvalues of $A(c)$ have the values*

$$\lambda_1^* = \cdots = \lambda_t^* < \lambda_{t+1}^* < \cdots < \lambda_{n-s}^*,$$

where $s = t(t-1)/2$.

It is clear that the numerical methods of §2 must now be modified since s of the rows of J have effectively been removed. Our goal is to obtain methods which are quadratically convergent even though s of the eigenvalues are not specified. Let us start by considering Method III, since in this case it is rather clear what should be done. Consider again (2.15), with second order terms neglected,

$$(Q^{(\nu)})^T A(c) Q^{(\nu)} = \Lambda^* + Y\Lambda^* - \Lambda^* Y, \tag{3.4}$$

which appears to represent $n(n+1)/2$ equations in the $n(n+1)/2$ unknowns c and Y. However $s = t(t-1)/2$ of the y_{ij}, namely those for which $1 \le i < j \le t$, are of no help in

solving (3.4), and may be removed from the equation, since they are multiplied by zero on the right-hand side of (3.4). Thus we see again that it is appropriate to specify only $n-s$ eigenvalues, and so we replace Λ^* in (3.4) by $\overline{\Lambda} = diag(\overline{\lambda}_i)$ where $\overline{\lambda}_i = \lambda_i^*$, $i = 1, \ldots, n-s$, and where the last s entries $\{\overline{\lambda}_i\}_{n-s+1}^n$ are free parameters. Equation (3.4) then has $n(n+1)/2$ unknowns, namely n parameters $\{c_k\}$, $n(n-1)/2 - s$ parameters $\{y_{ij}\}$, and s parameters $\{\overline{\lambda}_i\}_{n-s+1}^n$ To solve this equation we separate the computations of c and Y, as before. There are n equations defining c alone, namely

$$\sum_{k=1}^{n} \left((q_i^\nu)^T A_k q_i^\nu\right) c_k^{\nu+1} = \lambda_i^* - (q_i^\nu)^T A_0 q_i^\nu \qquad i = 1, \ldots, n-s \tag{3.5}$$

and

$$\sum_{k=1}^{n} \left((q_i^\nu)^T A_k q_j^\nu\right) c_k^{\nu+1} = -(q_i^\nu)^T A_0 q_j^\nu \qquad 1 \le i < j \le t. \tag{3.6}$$

The equations (3.6) were not previously imposed by Method III; they were not needed since we had assumed existence of a solution to an overdetermined problem. We denote the combined system (3.5), (3.6) by

$$K^{(\nu)} c^{\nu+1} = h^\nu. \tag{3.7}$$

Having thus defined $c^{\nu+1}$, the remaining unknowns $\{y_{ij}\}$ and $\{\overline{\lambda}_i\}$ are defined by

$$\overline{\lambda}_i = (q_i^\nu)^T A(c^{\nu+1}) q_i^\nu \qquad n-s < i \le n \tag{3.8}$$

$$y_{ij} = \frac{(q_i^\nu)^T A(c^{\nu+1}) q_j^\nu}{\overline{\lambda}_j - \overline{\lambda}_i} \qquad 1 \le i < j \le n, \quad j > t. \tag{3.9}$$

Finally, we set

$$y_{ij} = 0 \qquad 1 \le i < j \le t \tag{3.10}$$

since these parameters describe the rotation of the eigenvectors corresponding to the multiple eigenvalue, and therefore can be set to zero. The convergence analysis of the next section will show that zero is, in fact, the best choice for these parameters.

One difficulty remains: suppose that at the solution one of the eigenvalues that was not specified is actually multiple, *i.e.*, for some $j > n - s$, $\lambda_j(c^*) = \lambda_i(c^*)$, with $i = j \pm 1$. Then if the modified method is applied we will normally have that $|y_{ij}| \to \infty$ as $c^\nu \to c^*$, since $\{\overline{\lambda}_j - \overline{\lambda}_i\}$ will generally converge to zero. This can be avoided by introducing a tolerance parameter and setting y_{ij} to zero if $|\overline{\lambda}_j - \overline{\lambda}_i|$ drops below this parameter.

A more formal description of the modifed Method III will be given below. Let us first go back to Method I and modify it so that it solves Problem 2. To compute the new estimate of the parameters, the $n - s$ equations (3.5) for the distinct eigenvalues are combined with the s equations (3.6) for the multiple eigenvalue λ_1^*, to give a system identical to (3.7), except that Q refers to the computed eigenvectors $Q(c)$, rather than the approximations updated by Method III. With hindsight we can show that this is asymptotically equivalent to applying Newton's method to a reformulation of (2.1). Note that (3.5) and (3.6) can be written as

$$\begin{aligned} (Q_1^{(\nu)})^T A(c^{\nu+1}) Q_1^{(\nu)} &= \lambda_1^* I_t \\ (q_i^\nu)^T A(c^{\nu+1}) q_i^\nu &= \lambda_i^* \qquad i = t+1, \ldots, n-s, \end{aligned} \tag{3.11}$$

where $Q_1^{(\nu)} = [q_1^\nu, \ldots, q_t^\nu]$. Consider the Newton iteration on the nonlinear system

$$\begin{aligned} f_1(c) &= Q_1(c)^T A(c) Q_1(c) - \lambda_1^* I_t = 0 \\ \big(f_2(c)\big)_i &= q_i(c)^T A(c) q_i(c) - \lambda_i^* = 0 \qquad i = t+1, \ldots, n-s, \end{aligned} \tag{3.12}$$

where $Q_1(c) = [q_1(c), \ldots, q_t(c)]$. Note that f_1 represents $t(t+1)/2$ equations and f_2 consists of some of the components of f in (2.1). Differentiating f_1 with respect to c, one obtains

$$\dot{f}_1(c) = \dot{Q}_1(c)^T A(c) Q_1(c) + Q_1(c)^T \dot{A}(c) Q_1(c) + Q_1(c)^T A(c) \dot{Q}_1(c). \tag{3.13}$$

Differentiating $Q_1(c)^T Q_1(c) = I$ and using the relation $A(c) = Q(c)\Lambda(c)Q(c)^T$, (3.13) simplifies to become

$$\dot{f}_1(c) = Q_1(c)^T \dot{A}(c) Q_1(c) + B\Lambda_t(c) - \Lambda_t(c) B, \tag{3.14}$$

where $B = \dot{Q}_1(c)^T Q_1(c)$ and $\Lambda_t(c) = diag(\lambda_1(c), \ldots, \lambda_t(c))$. Assuming that B remains bounded, the last two terms in the right-hand side of (3.14) cancel in the limit since $\Lambda_t(c) \to \lambda_1^* I_t$. Thus from (2.4) and (3.14) we see that (3.11) is essentially a Newton step on (3.12). Although a true Newton step would include the last two terms on the right-hand side of (3.14), quadratic convergence is not impeded by dropping them. To prove quadratic convergence, one must take into account the lack of continuity of $Q_1(c)$, and this can be done by applying Rellich's theorem as discussed in §3.1. However, in the next section we will present a more direct proof.

Let us now modify Method II. Using the implementation of inverse iteration described in §3.1 we compute approximations $\{q_1^\nu, \ldots, q_{n-s}^\nu\}$ to the eigenvectors corresponding to the $n-s$ given eigenvalues $\{\lambda_i^*\}_1^{n-s}$. The new estimate of the parameters is obtained, as in Methods I and III, by solving (3.7), where the q_i^ν are the vectors obtained by inverse iteration.

Thus the Modified Methods I, II and III have the same form, with differences only in the way of computing the q_i^ν. We will obtain quadratic convergence for the three methods if we assume that the matrix K defined by (3.5)–(3.7), with $\{q_i^\nu\}$ replaced by a set of eigenvectors of $A(c^*)$, is nonsingular. Note that this condition is independent of the choice of the basis $Q_1(c^*)$, a much more satisfactory situation than in §3.1.

We now describe in detail the modified versions of Methods I, II and III designed for solving Problem 2.

Modified Method I

Choose c^0. Form $A(c^0)$ and compute its eigenvalues and eigenvectors.

For $\nu = 0, 1, 2, \ldots$

1. Stop if $\left[\sum_{i=1}^{n-s} \left(\lambda_i(c^\nu) - \lambda_i^*\right)^2\right]^{\frac{1}{2}}$ is sufficiently small.
2. Form $K^{(\nu)}$ and h^ν (see (3.5)–(3.7)), using $q_i^\nu = q_i(c^\nu)$. Compute $c^{\nu+1}$ by solving (3.7). Form $A(c^{\nu+1})$.
3. Find the eigenvalues $\{\lambda_i(c^{\nu+1})\}$ and eigenvectors $\{q_i(c^{\nu+1})\}$ of $A(c^{\nu+1})$. (Actually only the first $n-s$ eigenvalues and eigenvectors are needed.)

Modified Method II

Choose c^0, form $A(c^0)$ and compute its matrix of eigenvectors $Q(c^0)$. Set $Q^{(0)} \leftarrow Q(c^0)$. (As in Method I only the first $n-s$ eigenvectors are needed.)

For $\nu = 0, 1, 2, \ldots$

1. Stop if
$$\left\| (Q_{n-s}^{(\nu)})^T A(c^\nu) Q_{n-s}^{(\nu)} - \Lambda_{n-s}^* \right\|_F$$
is sufficiently small, where $Q_{n-s}^{(\nu)} = [q_1^\nu, \ldots, q_{n-s}^\nu]$ and $\Lambda_{n-s}^* = diag(\lambda_1^*, \ldots, \lambda_{n-s}^*)$.
2. Form $K^{(\nu)}$ and h^ν (see (3.5)–(3.7)), and compute $c^{\nu+1}$ by solving (3.7). Form $A(c^{\nu+1})$.
3. Compute the factorization
$$A(c^{\nu+1}) = UNU^T,$$
where U is orthogonal and N is tridiagonal. Solve
$$[N - \lambda_1^* I](U^T \Gamma) = U^T Q_1^{(\nu)}$$
for $\Gamma \in \mathbf{R}^{n \times t}$, where $Q_1^{(\nu)} = [q_1^\nu, \ldots, q_t^\nu]$, and compute the "$QR$" factorization
$$\Gamma = Q_1^{(\nu+1)} T,$$
where T is upper triangular (if necessary modify $Q_1^{(\nu)}$, as described in §3.1, to ensure that T has full column rank). Next solve
$$[N - \lambda_i^* I](U^T \gamma_i) = U^T q_i^\nu \qquad i = t+1, \ldots, n-s$$
and compute
$$q_i^{\nu+1} = \frac{\gamma_i}{\|\gamma_i\|} \qquad i = t+1, \ldots, n-s.$$

Modified Method III

Choose $c^{(0)}$ and set $Q^{(0)} \leftarrow Q(c^0)$. Choose a tolerance parameter $Neglig$.

For $\nu = 0, 1, 2, \ldots$

1. Stop if
$$\left\| (Q_{n-s}^{(\nu)})^T A(c^\nu) Q_{n-s}^{(\nu)} - \Lambda_{n-s}^* \right\|_F$$

is sufficiently small.

2. Form $K^{(\nu)}$ and h^{ν} (see (3.5)–(3.7)), and compute $c^{\nu+1}$ by solving (3.7). Form $A(c^{\nu+1})$.
3. For $1 \leq i < j \leq n$, compute

$$y_{ij} = \begin{cases} \dfrac{(q_i^{\nu})^T A(c^{\nu+1}) q_j^{\nu}}{\overline{\lambda}_j - \overline{\lambda}_i} & \text{if } |\overline{\lambda}_i - \overline{\lambda}_j| > Neglig \\ 0 & \text{otherwise} \end{cases}$$

where

$$\overline{\lambda}_i = \begin{cases} \lambda_i^* & \text{if } 1 \leq i \leq n-s \\ (q_i^{\nu})^T A(c^{\nu+1}) q_i^{\nu} & \text{otherwise.} \end{cases}$$

4. Compute $H = (I - \frac{1}{2}Y)Q^{(\nu)}$, factorize $(I + \frac{1}{2}Y)$ and use this to solve the n linear systems

$$(I + \frac{1}{2}Y)v_i = h_i \qquad i = 1, \ldots, n,$$

where h_i is the i^{th} column of H, and set

$$(Q^{(\nu+1)})^T = [v_1, \ldots, v_n].$$

To our knowledge, the modified methods are new. Indeed, we are not aware of any discussion of the correct *general* formulation of the inverse eigenvalue problem in the multiple eigenvalue case, where the dimension of the parameter space is chosen as in Problem 2. However, the principles on which these ideas rest are well known from the perturbation theory of multiple eigenvalues; see Davis and Kahan (1970), Friedland (1978) or Kato (1966). The dimension argument is essentially the same as the phenomenon known in quantum mechanics as the "crossing rule" of Von Neuman and Wigner (1929); see also Friedland,Robbin and Sylvester(1984).

The fact that the number of parameters must be increased in the multiple eigenvalue case is well known in the structural engineering literature; see Olhoff and Taylor (1983, p. 1147). The dimension argument has also been discussed in connection with the problem of communality (see Harman (1967)) and the educational testing problem (see Fletcher (1985)).

Since the problem of communality, described in §1.1, is an important special case that has received much attention, we will discuss it in more detail.

If the minimum rank is known to be $n - t$ then t of the eigenvalues λ_i^* are zero, but no other eigenvalues are given. Following our remarks at the beginning of this section, we know that the problem will be well posed in general if $n - s$ eigenvalues are specified, where, as before, $s = t(t-1)/2$. Thus the problem of communality will be well posed if $n - s = t$, or

$$n = \frac{t(t+1)}{2}. \tag{3.15}$$

This equation is solvable only for certain values of n. In particular, when $n = 6$ we can expect to be able to set $t = 3$, and the modified methods will be locally quadratically convergent. However, for $n = 7$ there is no value of t that will satisfy (3.15). When $t = 3$ the problem is underdetermined, and a natural course to take is to consider instead a minimization problem subject to the constraints $\lambda_1^* = \lambda_2^* = \lambda_3^* = 0$. The objective function could be, for example, the sum of the diagonal elements of the matrix. The formula (3.15) is well known and can be derived in several different ways; see Harman (1967, pp. 69-70) and Fletcher (1985, p. 502).

Harman describes various methods for solving the communality problem and other related problems, but none of them seems to be quadratically convergent. To our knowledge, the only algorithm for solving general minimal rank problems, which is known to be quadratically convergent, is that of Fletcher (1985). The spirit of Fletcher's algorithm is similar to that of our modified methods, since it makes the correct count of the number of equations and is also related to Newton's method. The algorithm is based on differentiating the block Cholesky factor corresponding to the null space of $A_0 + D$; it is actually derived for problems where $A_0 + D$ is constrained to be positive definite. The method does not coincide with any of our modified methods, and it does not seem to be possible to generalize it to handle other inverse eigenvalue problems.

There are various numerical methods, especially in the engineering literature, that are designed to handle optimization problems where multiple eigenvalues arise, either in the objective function or in the constraints (see for example Polak and Wardi (1982) and Cullum, Donath and Wolfe (1975)). Most of these are first-order methods, i.e. they

are not quadratically convergent. Choi and Haug (1981), however, give a quadratically convergent method for solving a specific design problem involving one double eigenvalue.

We complete this section with a discussion of local existence and uniqueness results for Problem 2. In the following theorem we will consider small perturbations which preserve eigenvalue multiplicities.

THEOREM 3.1. *Assume that c^* is a solution to Problem 2. Suppose that $K(c^*)$, defined by (3.5)–(3.7) replacing $\{q_i^\nu\}$ by any orthonormal set of eigenvectors of $A(c^*)$,is nonsingular. Then there exists $\epsilon > 0$ such that, for all $\{\mu_i^*\}_1^{n-s}$ satisfying*

$$\mu_1^* = \cdots = \mu_t^* < \mu_{t+1}^* < \cdots < \mu_{n-s}^*$$

and

$$|\mu_i^* - \lambda_i^*| \leq \epsilon \qquad 1 \leq i \leq n-s,$$

Problem 2, with $\{\lambda_i^*\}_1^{n-s}$ replaced by $\{\mu_i^*\}_1^{n-s}$, has a locally unique solution $d(\mu^*)$ with

$$\|c^* - d(\mu^*)\| = O(\epsilon).$$

PROOF: Solving the perturbed problem is equivalent to solving

$$F(d, X, M^*) = e^{-X} Q^T A(d) Q e^X - M^* = 0,$$

where Q is an orthogonal matrix of eigenvectors of $A(c^*)$, $M^* = diag(\mu_i^*)$ and X is a skew-symmetric matrix with the restriction that $x_{ij} = 0$ for $1 \leq i < j \leq t$. When $\mu_1^*, \ldots, \mu_{n-s}^*$ are fixed, this is an analytic system of $n(n+1)/2$ equations in $n(n+1)/2$ unknowns, since s of the $\{x_{ij}\}$ are set to zero and s of the $\{\mu_i^*\}$ are free. By expanding e^X and neglecting second-order terms, one sees that the Jacobian of F with respect to d, X and $\{\mu_i^*\}_{n-s+1}^n$ is nonsingular at $d = c^*$, $X = 0$ and $M^* = \Lambda^*$ if and only if $K(c^*)$ is nonsingular. The result therefore follows from the implicit function theorem; see Ortega and Rheinboldt (1970, p.128).

This theorem shows that Problem 2 is numerically well posed. If the perturbation does not preserve multiplicities, the perturbed problem is in general ill posed. In particular, some of the components x_{ij} may change from zero to arbitrarily large values.

§3.3 Convergence Analysis.

In this section we present convergence results for Methods I, II and III. For convenience, we first assume that all the eigenvalues

$$\lambda_1^* = \cdots = \lambda_t^* < \lambda_{t+1}^* < \cdots < \lambda_n^* \tag{3.16}$$

are specified, and analyze the methods in their unmodified form. When $t > 1$, this means that the problem is overdetermined as already explained, but this causes no difficulty for the convergence analysis since we assume existence of a solution. Afterwards, we explain how to adapt the proofs to apply to the modified methods, when only $\lambda_1^*, \ldots, \lambda_{n-s}^*$ are specified.

To start, we make the following assumptions.

Assumption 3.1

(i) There exists c^* such that $A(c^*)$ has eigenvalues given by (3.16).

(ii) The matrices $J(c^\nu)$ used in Method I satisfy $\limsup_{\nu\to\infty}\{\|J(c^\nu)^{-1}\|\} < \infty$, and the matrices $J^{(\nu)}$ of Methods II and III satisfy $\limsup_{\nu\to\infty}\{\|J^{(\nu)^{-1}}\|\} < \infty$.

When we analyze the modified methods, the second part, namely (ii), will be replaced by an assumption on the nonsingularity of $K(c^*)$.

We will now present several preliminary results that will be needed for the convergence proofs. Let $p_1, \ldots, p_t$ be any orthonormal set of eigenvectors of $A(c^*)$ corresponding to the multiple eigenvalue λ_1^*, and let $P_1 = [p_1 \ldots p_t]$. Denote the eigenprojection of $A(c^*)$ for λ_1^* by Π; we have

$$\Pi = P_1 P_1^T. \tag{3.17}$$

Let $P_2 = [p_{t+1} \dots p_n]$ be the matrix of orthonormal eigenvectors corresponding to λ^*_{t+1}, $\dots, \lambda^*_n$. Given any orthogonal matrix $Q = [Q_1\ Q_2]$, consider the problem of constructing an orthogonal matrix of eigenvectors of $A(c^*)$,

$$\widetilde{P} = [\widetilde{P}_1\ P_2], \tag{3.18}$$

which is, in some sense, close to Q. Note that there is freedom only in the way $\widetilde{P}_1$ is chosen; specifically $\widetilde{P}_1$ is of the form $P_1 B$, where B is an orthogonal matrix of order t. To find $\widetilde{P}_1$ we start by considering the matrix ΠQ_1 whose columns are eigenvectors for λ^*_1, but are not orthonormal. Then we form the "QR" factorization of ΠQ_1:

$$\Pi Q_1 = \widetilde{P}_1 R, \tag{3.19}$$

where R is a $t \times t$ nonsingular upper triangular matrix and $\widetilde{P}_1$ is an $n \times t$ matrix whose columns are orthonormal. Clearly the columns of $\widetilde{P}_1$ are eigenvectors of $A(c^*)$. Let us define the error matrices

$$\begin{aligned} E_1 &= Q_1 - \Pi Q_1 \\ E_2 &= Q_2 - P_2. \end{aligned} \tag{3.20}$$

LEMMA 3.1. *Let $P = [P_1\ P_2]$ be an orthogonal matrix of eigenvectors of $A(c^*)$. Then there exist constants $\epsilon > 0$ and $C > 0$ such that, for any orthogonal matrix $Q = [Q_1\ Q_2]$ with $\|E_1\| < \epsilon$, the matrix $\widetilde{P}$ defined by (3.17) and (3.19) satisfies*

$$Q^T \widetilde{P} = \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} \begin{bmatrix} \widetilde{P}_1 & P_2 \end{bmatrix} = \begin{bmatrix} I - F & -Q_1^T E_2 \\ E_2^T \widetilde{P}_1 & I - Q_2^T E_2 \end{bmatrix},$$

where $\|F\| \le C\|E_1\|^2$.

PROOF:

$$\begin{aligned} Q_1^T P_2 &= Q_1^T (Q_2 - E_2) = -Q_1^T E_2 \\ Q_2^T \widetilde{P}_1 &= (E_2 + P_2)^T \widetilde{P}_1 = E_2^T \widetilde{P}_1 \\ Q_2^T P_2 &= Q_2^T (Q_2 - E_2) = I - Q_2^T E_2. \end{aligned}$$

Let us now consider the remaining block. Since $\Pi\widetilde{P}_1 = \widetilde{P}_1$ we obtain from (3.19)

$$R = \widetilde{P}_1^T \Pi Q_1 = \widetilde{P}_1^T Q_1 \tag{3.21}$$

and

$$\begin{aligned} R^T R &= (\Pi Q_1)^T (\Pi Q_1) \\ &= (Q_1 - E_1)^T (Q_1 - E_1) \\ &= I - Q_1^T E_1 - E_1^T Q_1 + E_1^T E_1. \end{aligned} \tag{3.22}$$

Now

$$E_1^T Q_1 = E_1^T (E_1 + \Pi Q_1) = E_1^T E_1,$$

and therefore (3.22) gives

$$R^T R = I - E_1^T E_1.$$

By doing a Cholesky factorization to obtain R we see that, provided ϵ is small enough,

$$R = I - F \tag{3.23}$$

where $\|F\| \leq C\|E_1\|^2$. The result then follows from (3.21).

COROLLARY 3.1. *There exist constants $C > 0$, $\epsilon > 0$ such that, for any orthogonal matrix Q with $\|E\| = \|[E_1\ E_2]\| < \epsilon$, the skew-symmetric matrix X defined by*

$$e^X = Q^T \widetilde{P}$$

satisfies

$$\|X\| \leq C\|E\|$$

$$\|X_{11}\| \leq C\|E\|^2.$$

Here X_{11} is the $t \times t$ leading block of X.

PROOF: It follows immediately from Lemma 3.1 since $e^X = I + X + O(\|X\|^2)$.

These results show that, given an approximate matrix of eigenvectors Q, the eigenvectors at the solution may be chosen in the form $\widetilde{P}$, so that X, which describes to first order the rotation from the basis Q to the basis $\widetilde{P}$, is $O(\|E\|)$, and furthermore X_{11}, which describes the rotation of Q_1 to $\widetilde{P}_1$, is $O(\|E\|^2)$. This last fact will be needed for the analysis of Method III.

For the following Theorem we define the error matrix

$$E(c) = [E_1(c)\ E_2(c)] = [Q_1(c) - \Pi Q_1(c) \quad Q_2(c) - P_2].$$

Since $(I - \Pi)Q(c)$ is a Lipschitz continuous function of c (see for example Kato(1966)), $E(c)$ is also Lipschitz continuous and

$$\|E(c)\| \leq L\|c - c^*\| \tag{3.24}$$

holds for all c near c^*, where L is some constant.

THEOREM 3.1. *There exists $\epsilon > 0$ such that, if $\|c^0 - c^*\| \leq \epsilon$, the iterates $\{c^\nu\}$ of Method I converge quadratically to c^*.*

PROOF: Let $c = c^\nu, \bar{c} = c^{\nu+1}, Q = Q(c)$ and define $\widetilde{P}$ by (3.17)–(3.19) where $P = [P_1\ \ P_2]$ is any orthogonal matrix of eigenvectors of $A(c^*)$. Define X by $e^X = Q^T\widetilde{P}$. From Corollary 3.1, $\|X\| = 0(\|E\|)$. Thus from (3.24) we see that $\|E\|$ and $\|X\|$ can be made as small as we like by making $\|c - c^*\|$ small enough. Since $\widetilde{P}$ is a matrix of eigenvectors of $A(c^*)$, we have

$$e^X \Lambda^* e^{-X} = Q^T A(c^*) Q$$

and thus

$$\Lambda^* + X\Lambda^* - \Lambda^* X = Q^T A(c^*) Q + O(\|X\|^2). \tag{3.25}$$

The diagonal equations of (3.25) are

$$\lambda_i^* = q_i^T A(c^*) q_i + O(\|X\|^2) \qquad i = 1, \ldots, n,$$

and therefore

$$\lambda^* = J(c)c^* + b(c) + O(\|X\|^2). \tag{3.26}$$

The new iterate $\bar{c}$ is defined by

$$\lambda^* = J(c)\bar{c} + b(c)$$

and so it follows that

$$J(c)(\bar{c} - c^*) = O(\|X\|^2).$$

Finally, by the nonsingularity assumption (see (i) in Assumption (3.1)),

$$\begin{aligned} \|\bar{c} - c^*\| &= O(\|X\|^2) \\ &= O(\|E\|^2) \\ &= O(\|c - c^*\|^2). \end{aligned}$$

Note that there is no need to invoke Rellich's theorem, as is done in the proof of Nocedal and Overton (1983). Instead we have made use of the fact that the eigenprojection, and consequently $E(c)$, are Lipschitz continuous.

For the convergence proofs of Methods II and III we define the error matrix $E^{(\nu)}$ by

$$\begin{aligned} E^{(\nu)} &= [E_1^{(\nu)} \; E_2^{(\nu)}] \\ &= [(I - \Pi)Q_1^{(\nu)} \quad Q_2^{(\nu)} - P_2]. \end{aligned} \tag{3.27}$$

Since all eigenvalues, and the eigenvectors corresponding to distinct eigenvalues, are Lipschitz continuous functions in a neighborhood of c^*, there exist constants N_1 and N_2 such that

$$\|\lambda(c) - \lambda^*\| \leq N_1\|c - c^*\| \tag{3.28}$$

and

$$\|Q_2(c) - P_2\| \leq N_2\|c - c^*\| \tag{3.29}$$

hold for all c near c^*. Let us define

$$\delta = \min_{i \neq j,\ j>t} \{|\lambda_i^* - \lambda_j^*|\}.$$

Then for $i \neq k,\ k > t$, we have

$$\begin{aligned} |\lambda_k(c) - \lambda_i^*| &\geq |\lambda_k^* - \lambda_i^*| - |\lambda_k(c) - \lambda_k^*| \\ &\geq \delta - N_1\|c - c^*\|. \end{aligned} \tag{3.30}$$

Since the inverse iteration (2.13) is not defined when an eigenvalue $\lambda_i(c^\nu)$ coincides with a target eigenvalue, we will assume that, for all ν and for all $1 \leq i, j < n$, $\lambda_i(c^\nu) \neq \lambda_j^*$. This is not a restriction in practice, since it is known that inverse iteration will work even when a target eigenvalue coincides with an eigenvalue $\lambda_i(c^\nu)$ to machine accuracy; see Peters and Wilkinson (1971).

THEOREM 3.2. *There exists $\epsilon > 0$ such that, if $\|c^0 - c^*\| \leq \epsilon$, the iterates $\{c^\nu\}$ generated by Method II converge quadratically to c^*.*

PROOF: Let $Q = Q^{(\nu)}$, $\overline{Q} = Q^{(\nu+1)}$, $E = E^{(\nu)}$, $\overline{E} = E^{(\nu+1)}$ and define $\widetilde{P}$ by (3.17)–(3.19) with $Q = Q^{(\nu)}$. Let the skew-symmetric matrix X be defined by $e^X = Q^T\widetilde{P}$. By Corollary 3.1, $\|X\| = O(\|E\|)$. Following the same reasoning as in the proof of Theorem 3.1 we see that (3.25) holds, and its diagonal equations give

$$\lambda^* = Jc^* + b + O(\|X\|^2), \tag{3.31}$$

where $J = J^{(\nu)}$ and $b = b^\nu$. The new iterate $\bar{c}$ of Method II is defined by

$$\lambda^* = J\bar{c} + b,$$

and thus

$$J(\bar{c} - c^*) = O(\|X\|^2).$$

By the nonsingularity assumption on J

$$\|\bar{c} - c^*\| = O(\|E\|^2). \tag{3.32}$$

From (3.24) $\|E^{(0)}\| \leq L\|c^0 - c^*\|$. Let us assume inductively that

$$\|E\| = O(\|c - c^*\|), \tag{3.33}$$

and thus, if we can show that

$$\|\overline{E}\| = O(\|\bar{c} - c^*\|), \tag{3.34}$$

then interlacing (3.32) and (3.34) completes the proof. We analyse the two components of the error, $\overline{E}_1$ and $\overline{E}_2$ separately. Note that while $Q = [Q_1 \;\; q_{t+1} \ldots q_n]$ denotes the matrix iterate $Q^{(\nu)}$, $Q(c) = [Q_1(c) \;\; Q_2(c)] = [Q_1(c) \;\; q_{t+1}(c) \ldots q_n(c)]$ denotes the eigenvectors of $A(c)$.

Part I (bound on $\|\overline{E}_1\|$):

From (3.32) and (3.33) we have

$$\|\bar{c} - c^*\| = O(\|c - c^*\|^2). \tag{3.35}$$

Let

$$U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = Q(\bar{c})^T Q_1$$

so that

$$Q_1 = Q_1(\bar{c})U_1 + Q_2(\bar{c})U_2, \tag{3.36}$$

where U_1 is a $t \times t$ matrix and U_2 is an $(n-t) \times t$ matrix. Moreover, U_1 is invertible, as the following argument shows. We have

$$\begin{aligned} U_1^T U_1 &= Q_1^T Q_1(\bar{c}) Q_1(\bar{c})^T Q_1 \\ &= Q_1^T [I - Q_2(\bar{c}) Q_2(\bar{c})^T] Q_1 \\ &= I - \widehat{F}^T \widehat{F}, \end{aligned} \tag{3.37}$$

where

$$\widehat{F} = Q_2(\bar{c})^T Q_1.$$

Taking norms

$$\begin{aligned} \|\widehat{F}\| &\leq \|Q_2(\bar{c})^T (I - \Pi) Q_1\| + \|Q_2(\bar{c})^T \Pi Q_1\| \\ &\leq \|(I - \Pi) Q_1\| + \|Q_2(\bar{c}) - P_2\| \|\Pi Q_1\| + \|P_2^T \Pi Q_1\|. \end{aligned}$$

From (3.33), (3.29) and (3.35), and since the last term is zero, we conclude that

$$\|\widehat{F}\| = O(\|c - c^*\|). \tag{3.38}$$

Therefore $U_1^T U_1 = I + O(\|c - c^*\|^2)$. Provided $\|c - c^*\|$ is small enough $U_1^T U_1$, and hence U_1, are nonsingular. Let M_1 be a constant such that

$$\|U_1^{-1}\| \leq M_1. \tag{3.39}$$

The vectors corresponding to the multiple eigenvalue are updated by

$$\begin{aligned} [A(\bar{c}) - \lambda_1^* I] \Gamma &= Q_1 \\ \Gamma &= \overline{Q}_1 T. \end{aligned} \tag{3.40}$$

To simplify the analysis we will assume that T is nonsingular, *i.e.*, there is no need to replace columns of Q_1 by columns of the identity matrix (see the discussion in Section 3.1). From (3.36) and (3.40)

$$\begin{aligned} \Gamma &= [A(\bar{c}) - \lambda_1^* I]^{-1} Q_1 \\ &= Q(\bar{c}) [\Lambda(\bar{c}) - \lambda_1^* I]^{-1} Q(\bar{c})^T Q_1 \\ &= \begin{bmatrix} Q_1(\bar{c}) & Q_2(\bar{c}) \end{bmatrix} \begin{pmatrix} [\Lambda_1(\bar{c}) - \lambda_1^* I]^{-1} & 0 \\ 0 & [\Lambda_2(\bar{c}) - \lambda_1^* I]^{-1} \end{pmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \end{aligned} \tag{3.41}$$

where $\Lambda_1 = diag(\lambda_1, \ldots, \lambda_t)$ and $\Lambda_2 = diag(\lambda_{t+1}, \ldots, \lambda_n)$. Therefore,

$$\Gamma = Q_1(\bar{c})[\Lambda_1(\bar{c}) - \lambda_1^* I]^{-1} U_1 + Q_2(\bar{c})[\Lambda_2(\bar{c}) - \lambda_1^* I]^{-1} U_2. \tag{3.42}$$

From (3.40) it follows that $T^T T = \Gamma^T \Gamma$, and thus (3.42) gives

$$\begin{aligned} T^T T &= U_1^T [\Lambda_1(\bar{c}) - \lambda_1^* I]^{-2} U_1 + U_2^T [\Lambda_2(\bar{c}) - \lambda_1^* I]^{-2} U_2 \\ &= U_1^T [\Lambda_1(\bar{c}) - \lambda_1^* I]^{-1} (I + B^T B) [\Lambda_1(\bar{c}) - \lambda_1^* I]^{-1} U_1, \end{aligned} \tag{3.43}$$

where

$$B = [\Lambda_2(\bar{c}) - \lambda_1^* I]^{-1} U_2 U_1^{-1} [\Lambda_1(\bar{c}) - \lambda_1^* I].$$

Using (3.28), (3.39) and (3.30) we obtain

$$\|B\| \leq M_1 N_1 \|\bar{c} - c^*\| / (\delta - N_1 \|\bar{c} - c^*\|).$$

Provided $\|c - c^*\|$ is small enough, the denominator in this relation is bounded away from zero and we can write

$$\|B\| = O(\|\bar{c} - c^*\|).$$

Consider the Cholesky factorization of $(I + B^T B)$, *i.e.*, let $\widehat{R}$ be an upper triangular matrix such that

$$\widehat{R}^T \widehat{R} = (I + B^T B). \tag{3.44}$$

As with (3.23) it follows that

$$\widehat{R} = I + O(\|\bar{c} - c^*\|^2). \tag{3.45}$$

We now substitute (3.44) into (3.43) to obtain

$$T = \widehat{R}[\Lambda_1(\bar{c}) - \lambda_1^* I]^{-1} U_1. \tag{3.46}$$

Thus from (3.28) and (3.45)

$$\begin{aligned} \|T^{-1}\| &= \|U_1^{-1}[\Lambda_1(\overline{c}) - \lambda_1^* I]\widehat{R}^{-1}\| \\ &= O(\|\overline{c} - c^*\|). \end{aligned} \tag{3.47}$$

Now

$$\widehat{R} = TU_1^{-1}[\Lambda_1(\overline{c}) - \lambda_1^* I]. \tag{3.48}$$

We can now estimate the error E_1. From (3.40) $\overline{Q}_1 = \Gamma T^{-1}$, and thus from (3.42) and (3.48) we obtain

$$\begin{aligned} \|(I - \Pi)\overline{Q}_1\| \leq & \|(I - \Pi)Q_1(\overline{c})[\Lambda_1(\overline{c}) - \lambda_1^* I]^{-1}U_1T^{-1}\| \\ & + \|(I - \Pi)Q_2(\overline{c})[\Lambda_2(\overline{c}) - \lambda_1^* I]^{-1}U_2T^{-1}\| \\ \leq & \|(I - \Pi)Q_1(\overline{c})\|\|\widehat{R}^{-1}\| \\ & + \|\left(\Lambda_2(\overline{c}) - \lambda_1^* I\right)^{-1}\|\|T^{-1}\|. \end{aligned} \tag{3.49}$$

Finally, from (3.24), (3.45), (3.30) and (3.47) we conclude that

$$\|\overline{E}_1\| = \|(I - \Pi)\overline{Q}_1\| = O(\|\overline{c} - c^*\|). \tag{3.50}$$

Part II (bound on $\|\overline{E}_2\|$)

The vectors $\{q_i\}$ corresponding to the distinct eigenvalues are updated by

$$\begin{aligned} [A(\overline{c}) - \lambda_i^* I]\gamma_i &= q_i \\ \overline{q}_i = \gamma_i/\|\gamma_i\|, \qquad & i = t+1, \ldots, n. \end{aligned} \tag{3.51}$$

In what follows i denotes an integer with $t < i \leq n$. We can write

$$q_i = \sum_{k=1}^{n} \alpha_k q_k(\overline{c}), \tag{3.52}$$

for some scalars $\{\alpha_k\}$. From (3.51) we have

$$\gamma_i = \sum_{k=1}^{n} \frac{\alpha_k}{\lambda_k(\overline{c}) - \lambda_i^*} q_k(\overline{c}). \tag{3.53}$$

We will now show that if $\|c - c^*\|$ is small enough, $\alpha_i \neq 0$. The inequality

$$\|q_i - q_i(\overline{c})\| \leq \|q_i - q_i(c^*)\| + \|q_i(\overline{c}) - q_i(c^*)\|,$$

together with (3.33) and (3.29), gives

$$\|q_i - q_i(\overline{c})\| = O(\|c - c^*\|). \tag{3.54}$$

(Note that $q_i(c^*)$ is one of the columns of P_2). Thus

$$\begin{aligned}\sum_{\substack{k=1\\k\neq i}}^{n} \alpha_k^2 + (\alpha_i - 1)^2 &= \|q_i - q_i(\overline{c})\|^2 \\ &= O(\|c - c^*\|^2),\end{aligned}$$

and consequently

$$\alpha_i = 1 + O(\|c - c^*\|). \tag{3.55}$$

We will now show that $q_i(\overline{c})$ and $\overline{q}_i$ are nearly parallel. From (3.51) and (3.53) we have

$$\begin{aligned}q_i(\overline{c})^T \overline{q}_i &= \frac{\alpha_i}{\|\gamma_i\|\left(\lambda_i(\overline{c}) - \lambda_i^*\right)} \\ &= \left[\sum_{k\neq i} \frac{\alpha_k^2\left(\lambda_i(\overline{c}) - \lambda_i^*\right)^2}{\alpha_i^2\left(\lambda_k(\overline{c}) - \lambda_i^*\right)^2} + 1\right]^{-\frac{1}{2}} \\ &= 1 + O\left(\sum_{k\neq i} \frac{\alpha_k^2\left(\lambda_i(\overline{c}) - \lambda_i^*\right)^2}{\alpha_i^2\left(\lambda_k(\overline{c}) - \lambda_i^*\right)^2}\right)\end{aligned}$$

Using (3.28), (3.30) and (3.55) we obtain

$$q_i(\overline{c})^T \overline{q}_i = 1 + O(\|\overline{c} - c^*\|^2). \tag{3.56}$$

We can now estimate $\overline{E}_2$, componentwise. Equations (3.56) and (3.29) give

$$\begin{aligned}\|\overline{q}_i - q_i(c^*)\|^2 &= 2[1 - q_i(c^*)^T \overline{q}_i] \\ &= 2\left[1 - q_i(\overline{c})^T \overline{q}_i + \left(q_i(\overline{c}) - q_i(c^*)\right)^T \overline{q}_i\right] \\ &= O(\|\overline{c} - c^*\|).\end{aligned} \tag{3.57}$$

The proof is completed by combining (3.50) and (3.57).

THEOREM 3.3. *There exists $\epsilon > 0$ such that, if $\|E^{(0)}\| \leq \epsilon$, then the norms of the error matrices $\{\|E^{(\nu)}\|\}$ of Method III converge quadratically to zero.*

PROOF: Let $Q = Q^{(\nu)}$, $\overline{Q} = Q^{(\nu+1)}$, $E = E^{(\nu)}$, $\overline{E} = E^{(\nu+1)}$; define $\widetilde{P}$ by (3.17)–(3.19) with $Q = Q^{(\nu)}$, and define X by $e^X = Q^T \widetilde{P}$. As for Methods I and II we obtain (3.25), provided $\|E\|$ is small enough. Moreover, from Corollary 3.1 we have both

$$\begin{aligned} X &= O(\|E\|) \\ X_{11} &= O(\|E\|^2). \end{aligned} \tag{3.58}$$

The matrix Y and vector $\overline{c}$ of Method III are defined by

$$\Lambda^* + Y\Lambda^* - \Lambda^* Y = Q^T A(\overline{c}) Q \tag{3.59}$$

and $Y_{11} = 0$, i.e.

$$y_{ij} = 0 \qquad 1 \leq i < j \leq t.$$

Subtracting (3.59) from (3.25) we get

$$(X - Y)\Lambda^* - \Lambda^*(X - Y) = Q^T\big(A(c^*) - A(\overline{c})\big)Q + O(\|X\|^2). \tag{3.60}$$

The diagonal equations give

$$J(\overline{c} - c^*) = O(\|X\|^2),$$

where $J = J^{(\nu)}$. By the nonsingularity assumption and (3.58) we have

$$\|\overline{c} - c^*\| = O(\|E\|^2). \tag{3.61}$$

The off-diagonal equations of (3.60) give, for $i > j$, $j > t$

$$x_{ij} - y_{ij} = \frac{1}{\lambda_j^* - \lambda_i^*} q_i^T\big(A(c^*) - A(\overline{c})\big)q_j + O(\|X\|^2).$$

It follows that

$$|x_{ij} - y_{ij}| = O(\|E\|^2) \qquad t < i < j. \tag{3.62}$$

Since $X_{11} = O(\|E\|^2)$ and $Y_{11} = 0$ by definition of the algorithm, (3.62) holds also for $i < j \leq t$. Therefore

$$\|X - Y\| = O(\|E\|^2) \tag{3.63}$$

and consequently, from (3.58)

$$\|Y\| = O(\|E\|). \tag{3.64}$$

Let us now look at the updated matrix

$$\overline{Q} = Q(I + \tfrac{1}{2}Y)(I - \tfrac{1}{2}Y)^{-1}.$$

We have

$$\begin{aligned}
\overline{Q} - \widetilde{P} &= Q\big[(I + \tfrac{1}{2}Y)(I - \tfrac{1}{2}Y)^{-1} - e^X\big] \\
&= Q\big[(I + \tfrac{1}{2}Y) - \big(I + X + O(\|X\|^2)\big)(I - \tfrac{1}{2}Y)\big](I - \tfrac{1}{2}Y)^{-1} \\
&= Q\big[Y - X + O(\|XY\| + \|X\|^2)\big](I - \tfrac{1}{2}Y)^{-1}.
\end{aligned}$$

Thus from (3.63) and (3.64)

$$\|\overline{Q} - \widetilde{P}\| = O(\|E\|^2). \tag{3.65}$$

Since $(I - \Pi)\widetilde{P}_1 = 0$ we have

$$\begin{aligned}
\overline{E}_1 &= (I - \Pi)\overline{Q}_1 \\
&= (I - \Pi)(\overline{Q}_1 - \widetilde{P}_1 + \widetilde{P}_1) \\
&= O(\|E\|^2)
\end{aligned}$$

and

$$\begin{aligned}\overline{E}_2 &= \overline{Q}_2 - P_2 \\ &= O(\|E\|^2).\end{aligned}$$

The proof indicates clearly why it is appropriate to set $Y_{11} = 0$, namely to obtain (3.62). It also follows that Y_{11} can have any value satisfying $Y_{11} = O(\|E\|^2)$. It is easy to modify this proof so as to show that the parameters $\{c^\nu\}$ converge quadratically to c^*. However, we have analyzed the behavior of the matrices $\{Q^{(\nu)}\}$ because one can view Method III esssentially as a procedure for generating these matrices (the computation of $\{c^\nu\}$ being only an intermediate step). Moreover, we will now show that the matrices $Q^{(\nu)}$ of Method III converge to a limit, which is not the case for Methods I and II.

COROLLARY 3.2. *Suppose that* $\|E^{(0)}\| \leq \epsilon$, *where* ϵ *is given by Theorem 3.3. Then the matrices* $\{Q^{(\nu)}\}$ *generated by Method III converge quadratically to a limit* Q^*.

PROOF: Note that

$$\begin{aligned}\overline{Q} - Q &= Q\big((I + \tfrac{1}{2}Y)(I - \tfrac{1}{2}Y)^{-1} - I\big) \\ &= Q\big(Y + O(\|Y\|^2)\big) \\ &= O(\|E\|),\end{aligned} \tag{3.66}$$

where the last step follows from (3.64). By Theorem 3.3 $\{Q^{(\nu)}\}$ is a Cauchy sequence and therefore has a limit $Q^* = [Q_1^* \; Q_2^*]$. (Observe that Q_2^* is equal to P_2.) Moreover, by the quadratic convergence of $\{\|E^{(\nu)}\|\}$, and (3.66), we have

$$\begin{aligned}\overline{Q} - Q^* &= \sum_{k=\nu+1}^{\infty} \big(Q^{(k)} - Q^{(k+1)}\big) \\ &= O\big(\|E^{(\nu+1)}\|\big) \\ &= O\big(\|E^{(\nu)}\|^2\big).\end{aligned} \tag{3.67}$$

Also

$$\begin{aligned}E_1 &= (I - \Pi)Q_1 \\ &= (Q_1 - Q_1^*) + \big[(Q_1^* - \overline{Q}_1) + (\overline{Q}_1 - \tilde{P}_1) + (\tilde{P}_1 - \Pi Q_1)\big].\end{aligned}$$

Using (3.67), (3.65), (3.19) and (3.23) we have

$$E_1 = (Q_1 - Q_1^*) + O(\|E\|^2). \tag{3.68}$$

Since $Q_2^* = P_2$ we also have

$$E_2 = Q_2 - Q_2^*. \tag{3.69}$$

Combining (3.68) and (3.69)

$$\begin{aligned} E &= \left[(Q_1 - Q_1^*) \quad (Q_2 - Q_2^*)\right] + O(\|E\|^2) \\ &= (Q - Q^*) + O(\|E\|^2). \end{aligned}$$

Therefore

$$E = O(\|Q - Q^*\|),$$

which together with (3.67) completes the proof.

Now we show how to adapt the proofs so that they apply to the modified algorithms designed for solving Problem 2. We are given only the $n - s$ smallest eigenvalues

$$\lambda_1^* = \cdots = \lambda_t^* < \lambda_{t+1}^* < \cdots < \lambda_{n-s}^*. \tag{3.70}$$

We replace Assumption 3.1 by

ASSUMPTION 3.2:

(i) There exists c^* such that $A(c^*)$ has eigenvalues given by (3.70)

(ii) The matrix $K(c^*)$, defined by (3.5)-(3.7) using any orthonormal set of eigenvectors of $A(c^*)$, is nonsingular.

THEOREM 3.4. *There exists $\epsilon > 0$ such that, if $\|c^0 - c^*\| \leq \epsilon$, the iterates $\{c^\nu\}$ of Modified Method I converge quadratically to c^*.*

PROOF: It follows the proof of Theorem 3.1 very closely. The only difference is that the new iterate $\bar{c}$ is defined by (3.5)–(3.7), where q_i^ν refers to the computed eigenvector $q_i(c^\nu)$.

Using the first $n - s$ diagonal equations in (3.25) plus the equations corresponding to $1 \leq i < j \leq t$ we obtain

$$K(c)(\bar{c} - c^*) = O(\|X\|^2).$$

The rest of the proof follows as before.

THEOREM 3.5. *There exists $\epsilon > 0$ such that, if $\|c^0 - c^*\| \leq \epsilon$ the iterates $\{c^\nu\}$ of Modified Method II converge quadratically to c^*.*

PROOF: Again, this follows the proof of Theorem 3.2 very closely, differing just as described for Theorem 3.4. Now the q_i^ν refer to the vectors updated by inverse iteration.

THEOREM 3.6. *There exists $\epsilon > 0$ such that, if $\|E^{(0)}\| \leq \epsilon$, the norms of the error matrices $\{\|E^{(\nu)}\|\}$ of Modified Method III converge quadratically to zero.*

PROOF: As in the proofs of the two previous theorems, replacing the marix J by K allows us to obtain

$$\bar{c} - c^* = O(\|E\|^2).$$

For the second part we need to consider the unspecified eigenvalues. From (3.25) and (3.8)

$$\bar{\lambda}_i - \bar{\lambda}_i^* = q_i^T A(\bar{c} - c^*)q_i + O(\|E\|^2) \qquad n - s < i \leq n.$$

The off-diagonal equations not included in K give

$$x_{ij} - y_{ij} = \frac{1}{\bar{\lambda}_j - \bar{\lambda}_i} q_i^T A(c^* - \bar{c})q_j + O(\|E\|^2) \qquad i < j, \quad j > t.$$

It follows that $x_{ij} - y_{ij} = O(\|E\|^2)$, provided the unspecified eigenvalues at the solution are each distinct from all other eigenvalues. As explained before, this last condition can be removed by introducing a tolerance parameter *Neglig* (see step 3 of Modifed Method III).

§4. NUMERICAL RESULTS

We have tested Methods I, II III and IV on various types of problems. In our experience, Method IV almost always requires more iterations for convergence than the other methods, and also encounters difficulties more often. On the other hand, the local behaviour of Methods I, II and III is very similar, as is illustrated by the three examples we present below. The tests were made on VAX 11/780s, at the Courant Mathematics and Computing Laboratory and at Northwestern University. Double precision arithmetic was used throughout, i.e. approximately 14 decimal digits of accuracy. The eigenvalues and eigenvectors were computed using the EISPACK subroutines; see Smith *et al* (1967). The starting points were chosen close to the solution, so that few iterations were required for convergence. A line search (or a trust region strategy) would be essential to make the algorithms convergent in practical applications. However, we have not included these features and have concentrated on the local behavior of the methods. In particular, we were interested in verifying that quadratic convergence takes place in practice, in both the distinct and the multiple eigenvalue cases.

We programmed the Modified Methods I, II and III as described in § 3. The iterations were stopped when the residual, defined in Step 1 of each method, was less than 10^{-8}. The parameter *Neglig* required by Method III was set to 10^{-12}. For convenience, all vectors will be written as row-vectors. When specifying a symmetric matrix we will only write its lower triangular part.

Example 1 This is an additive inverse problem with distinct eigenvalues. Here $n = 8$,

$$A_0 = \begin{bmatrix} 0 & & & & & & & \\ 4 & 0 & & & & & & \\ -1 & -1 & 0 & & & & & \\ 1 & 2 & 3 & 0 & & & & \\ 1 & 1 & 1 & 1 & 0 & & & \\ 5 & 4 & 3 & 2 & 1 & 0 & & \\ -1 & -1 & -1 & -1 & -1 & -1 & 0 & \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 0 \end{bmatrix}, \quad A_k = e_k e_k^T \quad k = 1, \ldots, 8$$

$$\lambda^* = (10, 20, 30, 40, 50, 60, 70, 80)$$

$$c^0 = (10, 20, 30, 40, 50, 60, 70, 80)$$

$$c^* = (11.90788, 19.70552, 30.54550, 40.06266,$$

$$51.58714, 64.70213, 70.17068, 71.31850).$$

The following table displays the values of the residual, for each method.

Iteration	Method I	Method II	Method III
0	$6.40E+00$	$6.40E+00$	$6.40E+00$
1	$8.93E-01$	$1.51E+00$	$1.23E+00$
2	$1.03E-01$	$9.74E-02$	$1.45E-01$
3	$2.72E-03$	$1.97E-03$	$3.48E-03$
4	$2.32E-06$	$1.14E-06$	$2.58E-06$
5	$1.69E-12$	$4.04E-13$	$1.50E-12$

Example 2 We define a problem with multiple eigenvalues and $n = 8$. First consider the 8x5 matrix

$$V = \begin{bmatrix} 1 & -1 & -3 & -5 & -6 \\ 1 & 1 & -2 & -5 & -17 \\ 1 & -1 & -1 & 5 & 18 \\ 1 & 1 & 1 & 2 & 0 \\ 1 & -1 & 2 & 0 & 1 \\ 1 & 1 & 3 & 0 & -1 \\ 2.5 & .2 & .3 & .5 & .6 \\ 2 & -.2 & .3 & .5 & .8 \end{bmatrix},$$

and define $B = I + VV^T$. Now define the matrices $\{A_k\}$ from B as follows: let $A_0 = 0$ and for $k = 1, \ldots, n$

$$A_k = \sum_{j=1}^{k-1} b_{kj}(e_k e_j^T + e_j e_k^T) + b_{kk} e_k e_k^T.$$

Now consider $\widehat{c} = (1, 1, 1, 1, 1, 1, 1, 1)$; by construction,

$$A(\widehat{c}) = B = I + VV^T.$$

It follows that 1 is a multiple eigenvalue of $A(\widehat{c})$ with multiplicity 3. In fact, the eigenvalues of $A(\widehat{c})$ are

$$(1, 1, 1, 2.12075, 9.21887, 17.2814, 35.7082, 722.681).$$

Now let us choose the target eigenvalues λ^*. A suitable choice is

$$\lambda^* = (1, 1, 1, 2.1, 9.0)$$

i.e. specifying one eigenvalue of multiplicity 3 and 2 distinct eigenvalues. Then $t = 3, s = 3$ and $n - s = 5$, and the dimensions are properly chosen for the formulation of Problem 2. We could use $\widehat{c}$ as a starting point for the methods, but instead we choose

$$c^0 = (.99, .99, .99, .99, 1.01, 1.01, 1.01, 1.01).$$

The locally unique solution found by all three modified methods is

$$c^* = (.9833610, .9743705, .9753132, 1.054523,$$
$$.8554860, .9117770, .9283310, .8880013).$$

The following table displays the residual for the three methods.

Iteration	Method I	Method II	Method III
0	$2.09E-01$	$2.09E-01$	$2.09E-01$
1	$1.92E-01$	$2.26E-01$	$2.79E-01$
2	$2.04E-01$	$1.54E-01$	$1.99E-02$
3	$3.23E-02$	$2.03E-02$	$1.26E-02$
4	$7.11E-03$	$2.45E-03$	$2.67E-04$
5	$1.44E-04$	$2.19E-05$	$3.18E-07$
6	$7.89E-08$	$1.85E-09$	$3.55E-13$
7	$3.66E-14$		

Example 3 This is an additive inverse problem with multiple eigenvalues. Here $n = 6$,

$$A_0 = \begin{bmatrix} 0 & & & & & \\ 6.3 & 0 & & & & \\ -1 & -3.7 & 0 & & & \\ -2 & -6 & .3 & 0 & & \\ 1 & 3 & -1 & -2.7 & 0 & \\ 6 & 12 & -4 & 4.0 & 1.3 & 0 \end{bmatrix}, \quad A_k = e_k e_k^T \quad k = 1, \ldots, 6$$

$\lambda^* = (0, 0, 0)$

$c^0 = (3, 14, 3, 14, 1, 18)$

$c^* = (3.308477, 14.17183, 2.225671, 13.54877, .9512727, 17.67949).$

Note that the problem is well posed by specifying only one eigenvalue of multiplicity three. The residual values are given below.

Iteration	Method I	Method II	Method III
0	$2.47E-01$	$2.47E-01$	$2.47E-01$
1	$1.50E-01$	$1.48E-01$	$1.47E-01$
2	$1.43E-02$	$2.29E-02$	$2.58E-02$
3	$2.89E-04$	$5.71E-04$	$6.58E-04$
4	$9.63E-08$	$3.76E-07$	$4.97E-07$
5	$1.22E-14$	$1.86E-13$	$3.21E-13$

These examples, and our overall numerical experience with the three methods, indicate that quadratic convergence indeed occurs in practice, and that the three methods have very similar local behavior.

Acknowledgements. We are very grateful to Gene Golub who provided us with several references and made numerous helpful comments during the course of this work. We would also like to thank Olof Widlund for many helpful conversations and Robert Kohn for bringing the structural engineering literature to our attention.

REFERENCES

F.W. Biegler-König (1981), *A Newton iteration process for inverse eigenvalue problems*, Numer. Math. **37**, 349–354.

Z. Bohte (1967–68), *Numerical solution of the inverse algebraic eigenvalue problem*, Comp. J. **10**, 385–388.

G. Borg (1946), *Eine Umkehrung der Sturm-Liouvilleschen Eigenwertaufgabe*, Acta. Math. **78**, 1–96.

P. J. Brussard and P. W. Glaudemans (1977), "Shell model applications in nuclear spectroscopy", Elsevier.

K. K. Choi and E. J. Haug (1981), *A numerical method for optimization of structures with repeated eigenvalues*, in Optimization of Distributed Parameter Structures, vol I (Eds. E. J. Haug and J. Cea), 534–551, Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands.

J. Cullum, W.E. Donath and P. Wolfe (1975), *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Study, **3**, 35–55.

C. Davis and W. M. Kahan (1970), *The rotation of eigenvectors by a perturbation, III*, SIAM J. Numer. Anal. **7**, 1-46.

C. de Boor and G.H. Golub (1978), *The numerically stable reconstruction of a Jacobi matrix from spectral data*, Linear Algebra and its Appl. **21**, 245–260.

J. E. Dennis and R. B. Schnabel (1983), "Numerical Methods for Unconstrained Optimization and Nonlinear Equations", Prentice Hall.

A.C. Downing and A.S. Householder (1956), *Some inverse characteristic value problems*, J. Assoc. Comput. Mach. **3**, 203–207.

R. Fletcher (1985), *Semi-definite matrix constraints in optimization*, SIAM J. on Control and Optimization **23**, 493–513.

S. Friedland (1977), *Inverse eigenvalue problems*, Lin. Alg. Appl. **17**, 15–51.

S. Friedland (1979), *The reconstruction of a symmetric matrix from the spectral data*, J. of Math. Anal. and Applic. **71**, 412–422.

S. Friedland (1978), *Extremal eigenvalue problems*, Bulletin of Braz. Math. Soc. **9**, 13–40.

S. Friedland, J. W. Robbin and J. H. Sylvester (1984), *On the crossing rule*, Comm. Pure and Appl. Math. **37**, 19–37.

I. M. Gelfand and B.M. Levitan (1955), *On the determination of a differential equation from its spectral function*, Amer. Math. Soc. Transl. Series Z **1**, 253–304.

G. H. Golub and C. F. Van Loan (1983), "Matrix Computations", Johns Hopkins University Press.

O. Hald (1972), *On discrete and numerical Sturm-Liouville problems*, Ph.D. dissertation, Dept. of Math., New York University.

H. Harman (1967), "Modern Factor Analysis", The University of Chicago Press.

K. J. Holzinger and H. Harman (1941), "Factor Analysis", The University of Chicago Press.

T. Kato(1966), "Perturbation theory for linear operators", Springer-Verlag.

W. N. Kublanovskaja (1970), *On an approach to the solution of the inverse eigenvalue problem (in Russian),*, Zapiski naučnych seminarov Leningradskogo Otdelenija Matematičeskogo Instituta, in V.A.Steklova Akademii Nauk SSSR, 138–149.

P. Lancaster (1964-a), *Algorithms for lambda-matrices*, Numer. Math. **6**, 388–394.

P. Lancaster (1964-b), *On eigenvalues of matrices dependent on a parameter*, Numer. Math **6**, 377–387.

D. Q. Mayne and E. Polak (1982), *Algorithms for the design of control systems subject to singular value inequalities*, Math. Programming Study **18**, 112-134.

J. Nocedal and M. L. Overton (1983), *Numerical methods for solving inverse eigenvalue problems*, Lecture Notes in Mathematics No. 1005, (eds. V. Pereyra and A. Reinoza), 212–226, Springer Verlag.

N. Olhoff and J. E. Taylor (1983), *On structural optimization*, Journal of Applied Mechanics **50**, 1138–1151.

J. M. Ortega (1972), "Numerical Analysis", Academic Press.

J. M. Ortega and W.C. Rheinboldt (1970), "Iterative solution of nonlinear equations in several variables,", Academic Press.

M. R. Osborne (1971), *On the inverse eigenvalue problem for matrices and related problems for difference and differential equations*, Conference on Application of Numerical Analysis , Lecture Notes in Mathematics no. 228, 155–168, Springer-Verlag.

B. N. Parlett (1980), "The symmetric eigenvalue problem", Prentice Hall.

G. Peters and J. H. Wilkinson (1971), *The calculation of specified eigenvectors by inverse iteration*, in Handbook for Automatic Computation Vol. II: Linear Algebra (eds. J. H. Wilkinson and C. Reinsch), Springer Verlag.

J. Pliva and S. Toman (1966), *Multiplicity of solutions of the inverse secular problem*, J. Mol. Spectry. **21**, 362–371.

E. Polak and Y. Wardi (1982), *Nondifferentiable optimization algorithm for designing control systems having singular value inequalities*, Automatica **18**, 267-283.

F. Rellich (1969), "Perturbation theory of eigenvalue problems", Gordon and Breach, New York.

B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe , V. C. Klema and C. B. Moler, *Matrix Eigensystem Routines-EISPACK Guide*, Lecture Notes in Computer Science **6**, Springer Verlag.

J. von Neuman and E. Wigner (1929), *Uber das Verhalten von Eigenwerten bei Adiabatischen Prozessen*, Phys. Zschr. **30**, 467–470.